Robust Estimation

Tuesday, September 14, 2021 7:09 PM

Many examples of learning models from data:

Gaussians, Mixtures, ICA, topic models, dictionaries etc.

Efficient algorithms for them.

But what if not all the data is generated by the model? e.g. $(1-\varepsilon)$ is from Model
ε is arbitrary!

Even worse, <u>malicious adversary</u> replaces
ε fraction of data with points of their choice.

Can we still learn/estimate the model parameters?

Consider isotropic transformation or SVD.

- model has $\mathbb{E}(X) = \mu = 0$ (say)
  if εn points can be replaced $\mathbb{E}(X)$ can be ... anything
  even if 1 point is replaced!

What about singular/eigenvectors?

— What about singular / eigenvectors?

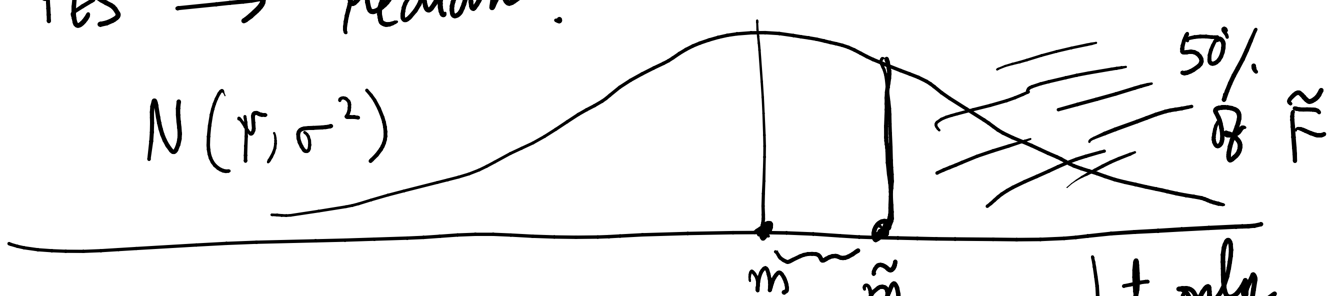
Say model has $v_1 = e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

if we add a point
far enough away along e_2, then $\tilde{v}_1 \to e_2$!

$\boxed{\text{So, mean, variance and low-degree sample moments are } \underline{\text{not}} \text{ robust estimates.}}$

Can we estimate a K-GMM with noise?
How about for $K=1$, i.e. a single Gaussian?
Let's even assume $\Sigma = I$, i.e. $N(\mu, I)$.

Can't do sample mean ....
How about in one dimension?
YES $\to$ Median!



but only

So $|m - \tilde{m}| \leq \varsigma$ $\qquad \frac{1}{2} - c \cdot |m - \tilde{m}|$ of F.

So $c\frac{|m-\tilde{m}|}{\sigma} \leq \varepsilon$ $\qquad \frac{1}{2} - c.\frac{|m-\tilde{m}|}{\sigma}$ of F.

$$|m-\tilde{m}| = O(\varepsilon)\cdot\sigma.$$

This is best possible "agnostic" estimate in 1-d.

Higher dim? median along each coordinate? $\times$

⋮

Robust Statistics $\quad$ Huber; Tukey. (~1960)

Tukey Ellipsoid: Smallest ellipsoid containing half the points.

Great estimator, but hard to compute!

[2016] For a large class of distributions including arbitrary Gaussians, we can robustly estimate mean, covariance up to information-theoretic limits.

information-theoretic limits.

$$\| \Sigma^{-\frac{1}{2}} (\hat{\mu} - \mu) \| \leq O\left(\varepsilon \sqrt{\log \tfrac{1}{\varepsilon}}\right).$$

Error does not grow with dimension!

Two simple algorithms:

① Iterative filtering

② recursive dimension halving.

Both based on following two ideas.

<u>Lemma</u>: Suppose Gaussian has $\Sigma = I$

If $\|\tilde{\Sigma}\|_2 \leq 1 + \varepsilon \Rightarrow \|\tilde{\mu} - \mu\|_2 = O\left(\varepsilon \sqrt{\log \tfrac{1}{\varepsilon}}\right)$

if noise is only additive, then $\text{—} = O(\varepsilon)$.

<u>Proof</u>: $G \cup B = S$

Good, Bad/Noise, Sample.

Assume true $\mu = 0$.

Noise.

$$\tilde{\mu} = (1-\varepsilon)\mu + \varepsilon \mu_B$$

Assume true $\mu = 0$.

then $\tilde{\mu} = \varepsilon \mu_B$.

$$\tilde{\Sigma} = (1-\varepsilon) I + \varepsilon \Sigma_B + (\varepsilon - \varepsilon^2) \mu_B \mu_B^T$$

Σ_G

(in additive model)

$\therefore$ for $V = \frac{\mu_B}{\|\mu_B\|}$

$$1 + \varepsilon \geq v^T \tilde{\Sigma} v^T \geq 1 - \varepsilon + (\varepsilon - \varepsilon^2)\|\mu_B\|^2$$

$$2\varepsilon \geq \frac{(\varepsilon - \varepsilon^2)}{\varepsilon^2} \|\mu\|^2$$

$$\Rightarrow \|\mu\| = O(\varepsilon).$$

With general noise $\|\mu\| = O\left(\varepsilon\sqrt{\log \frac{1}{\varepsilon}}\right)$.

Idea 2: Remove points so that $\|\tilde{\Sigma}\|_2$ is close to 1.

How?

Suppose $\exists v$: $v^T \tilde{\Sigma} v > 1 + C\varepsilon\sqrt{\log \frac{1}{\varepsilon}}$.

Lemma. $\Pr(X > t) \leq e^{-t^2/2}$

Lemma. $\Pr(X > t) \ldots$

$X \sim N(0,1)$

If something like this holds, then

$$v^T \mathbb{E}_B(XX^T)\, v = O\left(\log \frac{1}{\varepsilon}\right)$$

and since $\Sigma = I + \varepsilon \Sigma_B + (\varepsilon - \varepsilon^2)\mu_B \mu_B^T + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$

$$v^T \Sigma v \le 1 + \varepsilon \cdot v^T \mathbb{E}_B(XX^T)\, v + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$$
$$= 1 + O\left(\varepsilon \log \frac{1}{\varepsilon}\right).$$

~~Suppose~~ $\exists t: \quad \Pr(X > t+2) > C \cdot e^{-t^2/2}$

remove all points outside t.
At least $\frac{1}{2}$ the points removed are from B.

By the end $\quad \|\Sigma\|_2$ is small
and at most 2ε points removed.

Remove points: $|X| > C\sqrt{d}$.

Lemma. $\lambda_{min}(\Sigma) \ge 1 - \varepsilon$.

$\mathrm{Tr}(\Sigma) \le d(1 + O(\varepsilon))$

$$\mathrm{Tr}(\Sigma) \le d(1+O(\varepsilon))$$

$$\Rightarrow \quad \lambda_{d/2}(\Sigma) \le 1+\varepsilon.$$

<u>Proof.</u> $\Sigma = (1-\varepsilon)I + \varepsilon\Sigma_B + (\varepsilon-\varepsilon^2)\mu_B\mu_B^T$

$\lambda_{min}(\Sigma) = v^T\Sigma v \geq (1-\varepsilon)$.

$$\begin{aligned}\mathrm{Tr}(\Sigma) &\le (1-\varepsilon)d + \varepsilon\left(\tfrac{1}{|B|}\sum X_i X_i^T\right)\\ &\le (1-\varepsilon)d + O(\varepsilon)\cdot d\\ &= (1+O(\varepsilon))d.\end{aligned}$$

So in bottom $d/2$ eigenspace, sample mean is a good approximation of true mean.

Recurse on top $d/2$ eigenspace.

$\longrightarrow$ Remove X_i: $\|X\| > c.\sqrt{\dim}$.

SVD.

log d levels of recursion $\Rightarrow$ error: $O(\varepsilon\sqrt{\log d})$