EE290 Mathematics of Data Science Lecture 23 - 11/21/2019

Lecture 23: Robust estimation of a location parameter

Lecturer: Jiantao Jiao *Scribe: Nived Rajaraman*

We continue the discussion on Huber's work on 1-dimensional robust estimation [Hub68] in this lecture. While discussion in the previous class focused on his previous paper [Hub64] which dealt with a fairly restrictive robust estimation setting where the adversary is assumed to make symmetric perturbations to the true distribution (implying that the sample mean is asymptotically consistent), we discuss a significant extension of this work in [Hub68] developing a minimax optimal interval estimate for the mean of a 1-dimensional distribution under a wide range distributional perturbations. To lay the foundation, we first introduce his work on robust hypothesis testing.

1 Robust hypothesis testing

We first introduce the simple hypothesis testing problem. Here, a sample $X \in \mathcal{X}$ is assumed to come from one of two distributions $\mathcal{H}_0 : P_0(x)$ (null hypothesis) and $\mathcal{H}_1 : P_1(x)$ (alternate hypothesis) and the objective is to test if X came from P_0 or P_1.

Definition 1. *Given a sample X, a test $\phi : \mathcal{X} \to [0, 1]$ returns the probability of rejecting the null hypothesis.*

The Neyman Pearson lemma stipulates that the Likelihood ratio test (LRT) is the optimal test for this simple hypothesis testing problem. To quantify this statement, we first define what it means to be optimal.

Definition 2. *The level of a simple hypothesis test is defined as $\mathbb{E}_{X\sim P_0}[\phi(X)]$ and is the probability of incorrectly rejecting when the sample indeed comes from the null hypothesis.*

Definition 3. *The power of a test is defined as $\mathbb{E}_{X\sim P_1}[\phi(X)]$ and is the probability of correctly accepting the alternate when the sample indeed comes from the alternate hypothesis.*

A good test would reject with low probability when the sample comes from the null hypothesis (and have low level) and reject with high probability when the sample comes from the alternate hypothesis (and have high power). The Neyman Pearson test quantifies that the LRT is uniformly most powerful - for all level-α tests, it is the test with the highest power.

Definition 4. *The LRT is defined with the likelihood ratio as the testing function. That is, defining the test statistic $h(X) = \frac{P_1(X)}{P_0(X)}$, the test rejects when $h(X)$ is large,*

$$\phi(X) = \begin{cases} 1, & h(X) > c, \\ 0, & h(X) < c, . \\ \gamma, & h(X) = c \end{cases}$$

for some $\gamma \in [0, 1]$.

Remark The choice of γ in the LRT is to "top-up" the level of the test and make it exactly α, and is typically required for discrete hypothesis distributions.

For the sake of clarity, we note that for the setting where $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_0$ or P_1 the test statistic is

$$h(X_1^n) = \sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_0(X_i)}.$$

One can immediately observe that the vanilla LRT is not robust in Huber's contamination model. Indeed the adversary can arbitrarily corrupt the distribution to make some samples X_i have $P_0(X_i) \to 0$, leading the LRT to incorrectly reject $\mathcal{H}_0$ when the original samples indeed came from P_0. Huber proposed a truncated LRT for the robust testing problem. The test statistic here is,

$$\widetilde{\Pi}(X_1^n) = \begin{cases} h(X_1^n) & \text{if } c_2 \le h(X_1^n) \le c_1, \\ c_1 & \text{if } h(X_1^n) > c_1, \\ c_2 & \text{if } h(X_1^n) < c_1 \end{cases} . \tag{1}$$

While it may seem like an arbitrary choice, nevertheless we will later show that this corresponds to performing the likelihood ratio test for the least favorable priors in the testing problem.
We also remark that in the composite hypothesis testing problem, the level and power of the test are defined as the level of the test under the worst-case distributions in the null and alternate respectively. That is,

$$\begin{aligned} \text{Level of } \phi \ &: \ \sup_{P \in \mathcal{H}_0} \mathbb{E}_{X \sim P}[\phi(X)] \\ \text{Power of } \phi \ &: \ \inf_{P \in \mathcal{H}_1} \mathbb{E}_{X \sim P}[\phi(X)] \end{aligned}$$

To put Huber's work into context, we also need to discuss the family of perturbations to the true distribution he considers. To this end, for any $0 \le \epsilon_0, \epsilon_1, \delta_0, \delta_1 < 1$ define,

$$\begin{aligned} \mathcal{P}_0(\epsilon_0, \delta_0) &\triangleq \{Q \in M(\mathbb{R}) : \forall\ x,\ Q(X < x) \ge (1-\epsilon_0) P_0(X < x) - \delta_0\} \\ \mathcal{P}_1(\epsilon_1, \delta_1) &\triangleq \{Q \in M(\mathbb{R}) : \forall\ x,\ Q(X > x) \ge (1-\epsilon_1) P_1(X > x) - \delta_1\} \end{aligned}$$

where $M(\mathbb{R})$ is the family of all distributions on $\mathbb{R}$. Pictorially these are as shown in Figure 1.

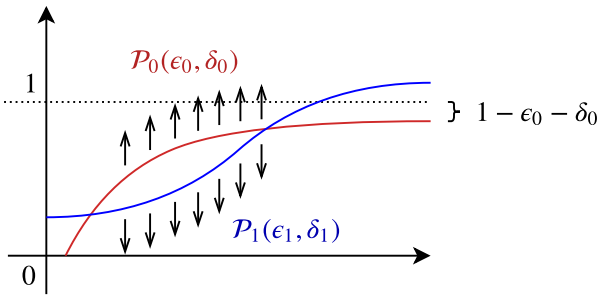

Figure 1: The families $\mathcal{P}_0$ and $\mathcal{P}_1$

For various choices of $(\epsilon_0, \delta_0, \epsilon_1, \delta_1)$ the families $\mathcal{P}_0$ and $\mathcal{P}_1$ include several commonly used divergences, including TV distance, Kolmogorov distance, Huber contamination and the Lévy Prokhorov metric among others. In particular, if the TV distance between distributions P and P_0 is $\le \epsilon$, then for all measurable sets A, $|P(A) - P_0(A)| \le \epsilon$. In particular this means that for all x, $|P(X < x) - P_0(X < x)| \le \epsilon$ which means that $P \in \mathcal{P}_0(0, \epsilon) \bigcap \mathcal{P}_1(0, \epsilon)$.

With this context, we formulate the fundamental testing problem. Huber considers a setting where the objective is to minimize the maximum of the type-I and type-II errors. That is,

$$\min_{\phi_j} \max_{j \in \{0,1\}} \sup_{Q_j \in \mathcal{P}_j} \mathbb{E}_{X \sim Q_j}[\phi_j(X)] \ \text{ such that } \phi_0 + \phi_1 = 1$$

Here ϕ_0 indicates the test which returns the probability of rejecting the null, and ϕ_1 its complement - which returns the probability of accepting the null.

Since we have a good handle on simple hypothesis testing, Huber proposed to find least favorable priors for the families $\mathcal{P}_0$ and $\mathcal{P}_1$. These are the distributions in $\mathcal{P}_0$ and $\mathcal{P}_1$ that are in some sense "closest" to each other and hardest to distinguish between. Intuitively, one would expect the distributions that achieve this property are those in Figure 2.

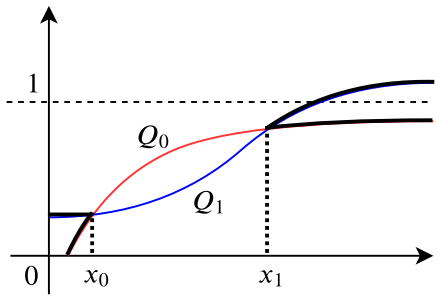

Figure 2: The least favorable priors Q_0 and Q_1

Indeed, Huber considers the distributions $Q_0(x)$ and $Q_1(x)$ that satisfy,

$$\forall x \in [x_0, x_1],\ Q_0(X < x) = (1-\epsilon_0)P_0(X < x) - \delta_0$$
$$\forall x \in [x_0, x_1],\ Q_0(X < x) = (1-\epsilon_1)P_1(X < x) + \epsilon_1 + \delta_1$$

In paritcular the densities for $j = 0, 1$ satisfy,

$$q_j(x) = ap_0(x) + bp_1(x),\ \forall x \geq x_0$$
$$q_j(x) = a'p_0(x) + b'p_1(x),\ \forall x \leq x_1$$
$$q_j(x) = (1-\epsilon_j)p_j(x),\ \forall x \in [x_1, x_1].$$

With this in place, observe that the optimal (in a uniformly most powerful sense) test to distinguish between Q_0 and Q_1 corresponds to the LRT with the test statistic as shown in eq. (1). Quantifying that this is indeed the least favorable prior comes from the following lemma.

Lemma 5. *For any $Q_0' \in \mathcal{P}_0$ and $Q_1' \in \mathcal{P}_1$, $Q_0'\widetilde{\Pi} < t) > Q_0(\widetilde{\Pi} < t)$ and $Q_1'\widetilde{\Pi} < t) > Q_1(\widetilde{\Pi} < t)$. That is,*

$$\widetilde{\Pi}\,|_{Q_0'} \preceq_{SD} \widetilde{\Pi}\,|_{Q_0}$$
$$\widetilde{\Pi}\,|_{Q_1'} \succeq_{SD} \widetilde{\Pi}\,|_{Q_1}\ .$$

where $\succeq_{SD}$ indicates stochastic dominance. In particular this means that any no other set of priors (i.e. hypotheses) can have higher type-I and type-II errors. Therefore, $\widetilde{\Pi}$ is the uniformly most powerful (UMP) test for robust testing.

This completes the picture for robust testing in 1-dimension. With this in place, we move back to the robust mean estimation setting.

2 Robust mean estimation

We consider the generalized location model (GLM), with samples $X_1, \cdots, X_n$ with $X_i - \theta \overset{\text{i.i.d.}}{\sim} \mathcal{L}_0$. This is more general than mean estimation, since we do not assume that the mean of $\mathcal{L}_0$ exists, and only care about estimation of θ. We show how to use results from the previous section on robust testing to construct optimal interval estimators for robust GLMs.

In a similar spirit as the robust testing framework introduced above, Huber considered the loss for any interval estimate $[b_1(X_1^n), b_2(X_1^n)]$ as being $\max\{P_\theta(b_1(X_1) > \theta), P_\theta(b_2(X_1) < \theta)\}$. The key idea behind his mean estimate is in constructing a test for $\mathcal{H}_0 : \theta < \theta_1$ vs. $\mathcal{H}_1 : \theta > \theta_2$ and inverting the test to get a confidence interval. For a more detailed exposition, and other related ideas, see §10 pp. 272-273 in Huber's book [HWI81].

We first consider testing between the composite hypothesis $\theta < \theta_1$ and $\theta > \theta_2$ (one should think of θ_1 as being negative in the following). Where it is clear from the context, we overload notation and write $X_1^n + \theta$ to indicate $\{X_1 + \theta, \cdots, X_n + \theta\}$. We assume that the test statistic $h(X_1^n + \theta)$ is monotone in θ, since the test would reject with higher probability if θ is larger. We consider the robust uniformly-most powerful test for these hypotheses discussed in the previous section,

$$\phi(X_1^n) = \begin{cases} 1, & h(X_1^n) > c \\ \gamma, & h(X_1^n) = c \\ 0, & h(X_1^n) < c \end{cases}.$$

In order to construct an interval estimate for θ, we simply "invert" the test after shifting. In particular, we define,

$$T^* = \sup\{\theta | h(X_1^n - \theta) > c\},$$
$$T^{**} = \inf\{\theta | h(X_1^n - \theta) < c\}.$$

That is, T^* is the largest value of θ, such that conditioned on the observation X_1^n, ϕ would reject the hypothesis $X_i - \theta \overset{\text{i.i.d.}}{\sim} \mathcal{L}_0$. Similarly, T^{**} is the smallest value of θ, such that conditioned on the observation X_1^n, ϕ would accept the hypothesis $X_i - \theta \overset{\text{i.i.d.}}{\sim} \mathcal{L}_0$. Note that by definition, for any θ, assuming continuity of h as a function of θ,

$$\{T^* > \theta\} = \{h(X_1^n - \theta) > c\}, \tag{2}$$
$$\{T^{**} \geq \theta\} = \{h(X_1^n - \theta) \geq c\}. \tag{3}$$

Also note that the estimator outputs,

$$T^\circ = \begin{cases} T^*, & \text{w.p. } 1 - \gamma, \\ T^{**}, & \text{w.p. } \gamma. \end{cases}$$

Then, the probability that the interval lies strictly above θ is,

$$P_\theta(T^\circ + \theta_1 > \theta) = (1-\gamma)P_\theta(T^* + \theta_1 > \theta) + \gamma P_\theta(T^{**} + \theta_1 > \theta) \overset{(i)}{\leq} \mathbb{E}_\theta[\phi(X_1^n + \theta_1 - \theta)] \overset{(ii)}{=} \alpha$$

where (i) follows from (2) and (3), and by noting that they are shift invariant, so $T^*(X_1^n + \delta) = T^*(X_1^n) + \delta$; (ii) follows from the fact that X_1^n comes from a GLM so,

$$\mathbb{E}_\theta[\phi(X_1^n + \theta_1 - \theta)] = \mathbb{E}_0[\phi(X_1^n + \theta_1)] = \mathbb{E}_{\theta_1}[\phi(X_1^n)] = \alpha$$

Similarly, we have that,

$$P_\theta(T^\circ + \theta_2 < \theta) = (1-\gamma)P_\theta(T^* + \theta_2 < \theta) + \gamma P_\theta(T^{**} + \theta_2 < \theta) \leq 1 - \mathbb{E}_\theta[\phi(X_1^n + \theta_2 - \theta)] = 1 - \beta$$

where β is the power of the test. Therefore, for any θ_1 and θ_2 we have constructed an interval estimate for θ (confidence interval) as $[T^\circ + \theta_1, T^\circ + \theta_2]$. In particular, since Huber's estimator minimizes $\max\{\alpha, 1-\beta\}$ as defined here, it is the optimal test to use for minimizing the error $\max\{P(b_1(X_1^n) < \theta), P(b_2(X_1^n) > \theta_2)\}$ for the interval estimate $[b_1(X_1^n), b_2(X_1^n)]$ for θ. Thus we have shown that the estimator is minimax optimal, since any estimator (for some θ_1 and θ_2) can be converted into test with level α and power β corresponding to the left and right ends of the interval, but cannot improve in $\max\{\alpha, 1-\beta\}$ over Huber's UMP robust test by Lemma 5.

References

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[Hub68] Peter J. Huber. Robust confidence limits. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 10(4):269–278, Dec 1968.

[HWI81] P.J. Huber, J. Wiley, and W. InterScience. *Robust statistics*. Wiley New York, 1981.