EE290 Mathematics of Data Science Lecture 22 - 11/19/2019

Lecture 22: Robust Location Estimation

Lecturer: Jiantao Jiao *Scribe: Vignesh Subramanian*

In this lecture, we get a historical perspective into the robust estimation problem and discuss Huber's work [1] for robust estimation of a location parameter.

The Huber loss function is given by,

$$\rho_{\mathsf{Huber}}(t) = \begin{cases} \frac{1}{2}t^2, & |t| \le k \\ k|t| - \frac{1}{2}k^2, & |t| > k \end{cases}. \tag{1}$$

Here k is a parameter and the idea behind the loss function is to penalize outliers (beyond k) linearly instead of quadratically. Figure 1 shows the Huber loss function for $k = 1$. In this lecture we will get an intuitive

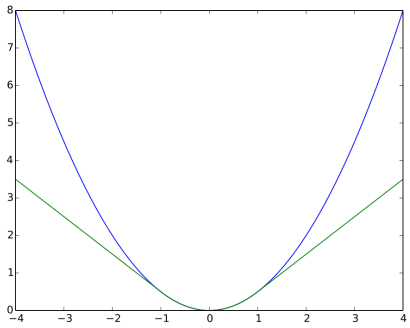

Figure 1: The green line plots the Huber-loss function for $k = 1$, and the blue line plots the quadratic function $\frac{1}{2}t^2$.

understanding for the reasons behind the particular form of this function, quadratic in interior, linear in exterior and convex and will see that this loss function is optimal for one dimensional robust mean estimation for Gaussian location model.

First we describe the problem setting.

1 Problem Setting

Suppose we observe $X_1, X_2, \ldots, X_n$ where $X_i - \mu \sim F \in \mathcal{F}$ are i.i.d. Here,

$$\mathcal{F} = \{F \mid F = (1-\epsilon)G + \epsilon H, H \in \mathcal{M}\}, \tag{2}$$

where $G \in \mathcal{M}$ is some fixed distribution function which is usually assumed to have *zero* mean, and $\mathcal{M}$ denotes the space of all probability measures. This describes the corruption model where the observed distribution is a convex combination of the true distribution $G(x)$ and an arbitrary corruption distribution H. It is a location model since we assume $X - \mu$ has distribution F where $\mu \in \mathbb{R}$ is unknown. The goal here is to estimate the parameter μ.

First we must determine how we evaluate estimators and in the paper, Huber restricted his attention to M-estimators of the form,

$$\hat{\mu} = \min_t \sum_{i=1}^{n} \rho(X_i - t).$$

As an example if $\rho(t) = \frac{1}{2}t^2$, then $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$, the empirical mean which is sensitive to outliers. To evaluate estimators Huber looks at asymptotics.

2 Asymptotics

Let $\psi(t) = \rho'(t)$. Then from first order condition of optimality, an optimizer T_n must satisfy,

$$\sum_{i=1}^n \psi(X_i - T_n) = 0. \tag{3}$$

Assume for now $\mu = 0$, and $\mathbb{E}_F[\psi(X)] = 0$. This means that for the population version of (3), $T_n = 0$ is a solution. We now assume that $T_n \to 0$ as $n \to \infty$, and we will provide a proof sketch showing T_n is asymptotically normal and compute its asymptotic variance.

From (3), using the first order approximation for the term $\psi(X_i - T_n)$ around the point X_i and using the mean-value theorem, for some $0 \leq \theta \leq 1$ we have,

$$\sum_{i=1}^n \psi(X_i) - T_n \sum_{i=1}^n \psi'(X_i - \theta T_n) = 0.$$

Rearranging we get,

$$\sqrt{n}T_n = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i)}{\frac{1}{n}\sum_{i=1}^n \psi'(X_i - \theta T_n)}.$$

Since we have $\mathbb{E}_F[\psi(X)] = 0$, the numerator by the Central Limit Theorem converges weakly to $N \sim \mathcal{N}(0, \mathbb{E}_F[\psi(X)^2])$. Further since we assumed $T_n \to 0$ as $n \to \infty$ then from the weak law of large numbers the denominator converges weakly to $\mathbb{E}_F[\psi'(X)]$. Thus we have,

$$\sqrt{n}(T_n - 0) \xrightarrow{w} \mathcal{N}\left(0, \frac{\mathbb{E}_F[\psi(X)^2]}{(\mathbb{E}_F[\psi'(X)])^2}\right).$$

One basic result for M-estimators is showing the maximum likelihood estimator achieves the smallest asymptotic variance among all M-estimators. We provide a proof below. Letting $f(x)$ denote the density function for F, we have

$$\begin{aligned}\mathbb{E}_F[\psi'(X)] &= \int_a^b f(x)\psi'(x)dx \\ &= f(x)\psi'(x)\Big|_a^b - \int_a^b \psi(x)f'(x)dx.\end{aligned}$$

If we assume that $f(a) = f(b) = 0$ then we have,

$$\mathbb{E}_F[\psi'(X)] = -\int_a^b \psi(x)f'(x)dx.$$

Thus,

$$\begin{aligned}\frac{\mathbb{E}_F[\psi(X)^2]}{(\mathbb{E}_F[\psi'(X)])^2} &= \frac{\int_a^b \psi(x)^2 f(x)dx}{\left(\int_a^b \psi(x)\left(\frac{f'(x)}{f(x)}\right)f(x)dx\right)^2} \\ &\geq \frac{1}{\left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx},\end{aligned}$$

where we used the Cauchy-Schwarz inequality. Observe that the RHS does not depend on ψ and the inequality is tight when $\psi(x) \propto -\frac{f'(x)}{f(x)}$ which results in $f(t) \propto e^{-\frac{\rho(t)}{A}}$ for some constant A. Thus minimizing $\rho(t)$ is equivalent to finding the maximum likelihood estimator. When $f(x)$ is a Gaussian density function, then ρ is the squared-loss function and the optimizer T_n is the empirical mean.

3 Two player game and Huber's Theorem

Consider a two player game with payoff function given by $-V(\psi, F)$. Here ψ is the action chosen by the statistician to maximize the payoff (minimize the asymptotic variance) and F is chosen by the adversary to minimize the payoff (maximize the asymptotic variance).

Theorem 1. *Assume G is symmetric around 0, log-concave with density function $g(x)$ with convex support. Define*

$$\mathcal{F}_{\mathsf{S}} = \{F \mid F = (1-\epsilon)G + \epsilon H, H \text{ symmetric around } 0\} \tag{4}$$

The two-player game under the assumptions describe above has a saddle point (ψ_0, F_0) i.e.,

$$\sup_{F \in \mathcal{F}_{\mathsf{S}}} V(\psi_0, F) = V(\psi_0, F_0) = \inf_{\psi} V(\psi, F_0).$$

First we describe the form of $f_0(x)$ which is the density function of F_0. Let $[t_0, t_1]$ be the interval where $\left|\frac{g'(x)}{g(x)}\right| \leq k$. We know that this interval exists since $g(x)$ is log-concave with convex support. Here k is the solution to the equation

$$\frac{1}{1-\epsilon} = \int_{t_0}^{t_1} g(t)dt + \frac{g(t_0) + g(t_1)}{k}. \tag{5}$$

Then,

$$f_0(t) = \begin{cases} (1-\epsilon)g(t_0)e^{k(t-t_0)}, & t \leq t_0 \\ (1-\epsilon)g(t), & t_0 < t < t_1 \\ (1-\epsilon)g(t_1)e^{-k(t-t_1)}, & t \geq t_1 \end{cases} \tag{6}$$

$$\psi_0(t) = -\frac{f_0'(t)}{f_0(t)}. \tag{7}$$

Before we look at the proof of this theorem we look at an example.

Example 2. Let $g(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. Then $-t_0 = t_1 = k$. We can solve either by binary search or line search for k using the equation,

$$\frac{1}{1-\epsilon} = \int_{-k}^{k} g(t)dt + \frac{2g(k)}{k}.$$

The optimal loss function to use in this case is the Huber loss function given by (1).

$$\rho_{\mathsf{Huber}}(t) = \begin{cases} \frac{1}{2}t^2, & |t| \leq k \\ k|t| - \frac{1}{2}k^2, & |t| > k \end{cases}.$$

Note that for a generic distribution $g(t)$ the dependence of t_0 and t_1 on k can be highly non-linear and it is not easy to solve for k using (5).
Next we look at the proof for Theorem 1.

Proof First we verify that the distribution H determined by F_0 and G is indeed a distribution i.e. its density function $h(t)$ is non-negative and integrates to one. We have,

$$\epsilon h_0(t) = \begin{cases} (1-\epsilon)(g(t_0)e^{k(t-t_0)} - g(t)), & t \leq t_0 \\ 0, & t_0 < t < t_1 \\ (1-\epsilon)(g(t_1)e^{-k(t-t_1)} - g(t)), & t \geq t_1 \end{cases}. \tag{8}$$

Since $g(t)$ and $f_0(t)$ integrate to one, $h(t)$ integrates to one. To show non-negativity of $h(t)$ we use the fact that $g(t)$ is log-concave, which implies $-\log(g(t))$ is a convex function. For any $t \leq t_0$,

$$\begin{aligned} -\log(g(t)) &\geq -\log(g(t_0)) - k(t-t_0), \\ \Rightarrow g(t) &\leq g(t_0)e^{k(t-t_0)}. \end{aligned}$$

where we used the the facts $\frac{g'(t_0)}{g(t_0)} = k$ and $(\log(g(t))' = \frac{g'(t)}{g(t)}$. The proof for the case $t \geq t_1$ follows via a similar argument. Next we need to show that $V(\psi_0, F_0)$ is a saddle point.

We have,

$$V(\psi_0, F_0) = \inf_{\psi} V(\psi, F_0),$$

because for given F_0, ψ_0 was optimal and resulted in the optimizer being the maximum likelihood estimator as discussed in Section 2. Next we show that,

$$V(\psi_0, F_0) = \sup_{F \in \mathcal{F}_{\mathsf{S}}} V(\psi_0, F).$$

For any $F \in \mathcal{F}_{\mathsf{S}}$ we have

$$V(\psi_0, F) = \frac{\mathbb{E}_F[\psi_0(X)^2]}{(\mathbb{E}_F[\psi_0'(X)])^2}.$$

We can rewrite the numerator as,

$$\begin{aligned} \mathbb{E}_F[\psi_0(X)^2] &= (1-\epsilon)\mathbb{E}_G[\psi_0(X)^2] + \epsilon\mathbb{E}_H[\psi_0(X)^2] \\ &\leq (1-\epsilon)\mathbb{E}_G[\psi_0(X)^2] + \epsilon k^2, \end{aligned}$$

where we upper $\mathbb{E}_H[\psi_0(X)^2]$ using $\psi_0(t) = -\frac{f_0'(t)}{f(t)}$ and the form of $f_0(t)$ from (6) which results in $|\psi(t)| = k$ for $t \leq t_0$ or $t \geq t_1$ and $|\psi(t)| = \left|\frac{g'(t)}{g(t)}\right| \leq k$ for $t_0 < t < t_1$. Note that $f_0(t)$ results in $h_0(t) = 0$ for $t_0 < t < t_1$ and thus maximizes the numerator.

Similarly the denominator can be written as,

$$\begin{aligned} (\mathbb{E}_F[\psi_0'(X)])^2 &= ((1-\epsilon)\mathbb{E}_G([\psi_0'(X)]) + \epsilon\mathbb{E}_H([\psi_0'(X)]))^2 \\ &\geq ((1-\epsilon)\mathbb{E}_G([\psi_0'(X)]))^2, \end{aligned}$$

where we used the fact that $\psi_0' \geq 0$ pointwise and $\psi_0'(t) = 0$ for $t \leq t_0$ or $t \geq t_1$.

Again note that $f_0(t)$ results in $h_0(t) = 0$ for $t_0 < t < t_1$ and minimizes the denominator. Thus F_0 is the maximizer of $V(\psi_0, F)$ among all $F \in \mathcal{F}_{\mathsf{S}}$. □

4 Summary

There were several criticisms of Huber's work including those on the assumptions that G and H are symmetric, and the requirement that ϵ be known in order to compute the Huber loss. Further in higher dimensions the breakdown point scales as $\frac{1}{1+d}$ which is undesirable. (From Wikipedia: Intuitively, the breakdown point of an estimator is the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle before giving an incorrect (e.g., arbitrarily large) result). In a subsequent paper Huber removes the assumptions that G, H are symmetric and shows that the Huber M-estimator is exactly minimax for coverage probability in robust location estimation for Gaussian models.

References

[1] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, Mar. 1964. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177703732