Theoretical Statistics. Lecture 15.

Peter Bartlett

M-Estimators.

Consistency of M-Estimators.

Nonparametric maximum likelihood.

M-estimators

Goal: estimate a parameter θ of the distribution P of observations $X_1, \ldots, X_n$.

Define a criterion $\theta \mapsto M_n(\theta)$ in terms of functions $m_\theta : \mathcal{X} \to \mathbb{R}$,

$$M_n(\theta) = P_n m_\theta.$$

The estimator $\hat{\theta} = \arg\max_{\theta \in \Theta} M_n(\theta)$ is called an **M-estimator** (M for maximum).

Example:
maximum likelihood uses

$$m_\theta(x) = \log p_\theta(x).$$

Z-estimators

Can maximize by setting derivatives to zero:

$$\Psi_n(\theta) = P_n \psi_\theta = 0.$$

These are **estimating equations**. van der Vaart calls this a **Z-estimator** (Z for zero), but it's often called an M-estimator (even if there's no maximization).

Example:
maximum likelihood:

$$\psi_\theta(x) = \nabla_\theta \log p_\theta(x).$$

M-estimators and Z-estimators

Of course, sometimes we cannot transform an M-estimator into a Z-estimator. Example: $p_\theta =$uniform on $[0, \theta]$ is not differentiable in θ, and there is no natural Z-estimator. The M-estimator chooses

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta P_n m_\theta \\ &= \arg\max_\theta P_n \log \frac{1\left[\cdot \in [0, \theta]\right]}{\theta} \\ &= \max_i X_i.\end{aligned}$$

M-estimators and Z-estimators: Examples

Mean:

$$m_\theta(x) = -(x-\theta)^2.$$

$$\psi_\theta(x) = (x-\theta).$$

Median:

$$m_\theta(x) = -|x-\theta|.$$

$$\psi_\theta(x) = \mathrm{sign}(x-\theta).$$

M-estimators and Z-estimators: Examples

Huber: [PICTURE]

$$m_\theta(x) = r_k(x-\theta)$$

$$r_k(x) = \begin{cases} \frac{1}{2}k^2 - k(x+k) & \text{if } x < -k, \\ \frac{1}{2}x^2 & \text{if } |x| \le k, \\ \frac{1}{2}k^2 + k(x-k) & \text{if } x > k. \end{cases}$$

$$\psi_\theta(x) = [x-\theta]_{-k}^{k}$$

$$[x]_{-k}^{k} = \begin{cases} -k & \text{if } x < -k, \\ x & \text{if } |x| \le k, \\ k & \text{if } x > k. \end{cases}$$

These are all location estimators: $m_\theta(x) = m(x-\theta)$, $\psi_\theta(x) = \psi(x-\theta)$.

Consistency of M-estimators and Z-estimators

We want to show that $\hat{\theta} \xrightarrow{P} \theta_0$, where $\hat{\theta}$ approximately maximizes $M_n(\theta) = P_n m_\theta$ and θ_0 maximizes $M(\theta) = P m_\theta$. We use a ULLN.

Theorem: Suppose that

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$,
2. For all $\epsilon > 0$, $\sup \{M(\theta) : d(\theta, \theta_0) \geq \epsilon\} < M(\theta_0)$, and
3. $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$.

Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

(2) is an identifiability condition: approximately maximizing $M(\theta)$ unambiguously specifies θ_0. It suffices if there is a unique maximizer, Θ is compact, and M is continuous.

Proof

From (2), for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$
\begin{aligned}
&\Pr(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \\
&\leq \Pr(M(\theta_0) - M(\hat{\theta}_n) \geq \delta) \\
&= \Pr(M(\theta_0) - M_n(\theta_0) + M_n(\theta_0) - M_n(\hat{\theta}_n) + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \geq \delta) \\
&\leq \Pr(M(\theta_0) - M_n(\theta_0) \geq \delta/3) + \Pr(M_n(\theta_0) - M_n(\hat{\theta}_n) \geq \delta/3) \\
&\qquad + \Pr(M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \geq \delta/3).
\end{aligned}
$$

Then (1) implies the first and third probabilities go to zero, and (3) implies the second probability goes to zero.

Consistency of M-estimators and Z-estimators

Same thing for Z-estimators: Finding $\hat{\theta}$ that is an approximate zero of $\Psi_n(\theta) = P_n \psi_\theta$ leads to θ_0, which is the unique zero of $\Psi(\theta) = P\psi_\theta$.

Theorem: Suppose that

1. $\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{P} 0$,
2. For all $\epsilon > 0$, $\inf \{\|\Psi(\theta)\| : d(\theta, \theta_0) \geq \epsilon\} > 0 = \|\Psi(\theta_0)\|$, and
3. $\Psi_n(\hat{\theta}_n) = o_P(1)$.

Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof: Choosing $M_n(\theta) = -\|\Psi_n(\theta)\|$ and $M(\theta) = -\|\Psi(\theta)\|$ in the previous theorem implies the result.

Example: Sample median

Sample median $\hat{\theta}_n$ is the zero of

$$P_n \psi_\theta(X) = P_n \operatorname{sign}(X - \theta).$$

Suppose that P is continuous and positive around the median, and check the conditions:

1. The class $\{x \mapsto \operatorname{sign}(x - \theta) : \theta \in \mathbb{R}\}$ is Glivenko-Cantelli.
2. The population median is unique, so for all $\epsilon > 0$,

$$P(X < \theta_0 - \epsilon) < \frac{1}{2} < P(X < \theta_0 + \epsilon).$$

3. The sample median always has $|P_n \operatorname{sign}(X - \hat{\theta}_n)| = 0$.

ULLN and M-estimators

Notice the ULLN condition:
$\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \stackrel{P}{\to} 0.$

Typically, this requires the empirical process $\theta \mapsto P_n m_\theta$ to be totally bounded. This can be problematic if m_θ is unbounded. For instance:
Mean: $m_\theta(x) = -(x-\theta)^2$,
Median: $m_\theta(x) = -|x-\theta|$.

We can get around the problem by restricting to a compact set where most of the mass of P lies, and showing that this does not affect the asymptotics. In that case, we can also restrict θ to an appropriate compact subset.

Non-parametric maximum likelihood

Estimate P on $\mathcal{X}$. Suppose it has a density

$$p_0 = \frac{dP}{d\mu} \in \mathcal{P},$$

where $\mathcal{P}$ is a family of densities. Define the maximum likelihood estimate

$$\hat{p}_n = \arg\max_{p \in \mathcal{P}} P_n \log p.$$

We'll show conditions for which $\hat{p}_n$ is **Hellinger consistent**, that is, $h(\hat{p}_n, p_0) \stackrel{as}{\to} 0$, where h is the Hellinger distance:

$$h(p, q) = \left(\frac{1}{2} \int \left(p^{1/2} - q^{1/2} \right)^2 d\mu \right)^{1/2}.$$

[The $1/2$ ensures $0 \leq h(p,q) \leq 1$.]

Hellinger distance

We have

$$h(p,q)^2 = \frac{1}{2}\int \left(p^{1/2} - q^{1/2}\right)^2 \, d\mu$$
$$= \frac{1}{2}\int \left(p + q - 2p^{1/2}q^{1/2}\right) \, d\mu$$
$$= 1 - \int p^{1/2}q^{1/2} \, d\mu.$$

This latter integral is called the Hellinger affinity. Expressing h in this form can simplify its calculation for product densities. Notice that, by Cauchy-Schwartz,

$$\int p^{1/2}q^{1/2} \, d\mu \leq \int p \, d\mu \int q \, d\mu = 1,$$

so $h(p,q) \in [0,1]$.

Non-parametric maximum likelihood

The Kullback-Leibler divergence between p and q is

$$d_{KL}(p,q) = \int \log \frac{q}{p} q \; d\mu.$$

Clearly, $d_{KL}(p,p) = 0$. Also, since $-\log(\cdot)$ is convex,

$$d_{KL}(p,q) = -\int \log \frac{p}{q} q \; d\mu \geq -\log \left(\int \frac{p}{q} q \; d\mu \right) = 0.$$

Non-parametric maximum likelihood

Relating KL-divergence to a ULLN:

$$
\begin{aligned}
d_{KL}(\hat{p}_n, p_0) &= \int \log \frac{p_0}{\hat{p}_n} \;\; p_0 \;\; d\mu \\
&\leq \int \log \frac{p_0}{\hat{p}_n} \;\; p_0 \;\; d\mu - P_n \log \frac{p_0}{\hat{p}_n} \\
&= P \log \frac{p_0}{\hat{p}_n} - P_n \log \frac{p_0}{\hat{p}_n} \\
&\leq \|P - P_n\|_G,
\end{aligned}
$$

where the first inequality follows from the fact that $\hat{p}_n$ maximizes $P_n \log p$

over $p \in \mathcal{P}$, and the class G is defined as

$$G = \left\{ 1[p_0 > 0] \log \frac{p_0}{p} : p \in \mathcal{P} \right\}.$$

Non-parametric maximum likelihood

One problem here is that $\log(p_0/p)$ is unbounded, since p can be zero. We'll take a different approach: For any $p \in \mathcal{P}$, consider the mixture

$$\tilde{p} = \frac{p + p_0}{2}.$$

If the class $\mathcal{P}$ is convex and $\hat{p}_n, p_0 \in \mathcal{P}$, this mixture has $P_n \log \tilde{p} \leq P_n \log \hat{p}_n$. This is behind the following lemma.

Lemma: Define

$$\tilde{p}_n = \frac{\hat{p}_n + p_0}{2}.$$

If $\mathcal{P}$ is convex,

$$h(\hat{p}_n, p_0)^2 \leq \int \frac{\hat{p}_n}{\tilde{p}_n} \, d(P_n - P).$$

Non-parametric maximum likelihood

Theorem: For a convex class $\mathcal{P}$ of densities, if P has density $p_0 \in \mathcal{P}$ and $\hat{p}_n$ maximizes likelihood over $\mathcal{P}$, we have

$$h(\hat{p}_n, p_0)^2 \leq \|P - P_n\|_G,$$

where

$$G = \left\{ \frac{2p}{p + p_0} : p \in \mathcal{P} \right\}.$$

Notice that functions in G are bounded between 0 and 2.

Non-parametric maximum likelihood: Example

Lemma: Suppose $\mathcal{P}$ is a set of densities on a compact subset $\mathcal{X}$ of $\mathbb{R}^d$. Fix a norm $\|\cdot\|$ on $\mathbb{R}^d$. Suppose that, for all $p \in \mathcal{P}$,

$$\left|\frac{p(x)}{p(y)} - 1\right| \leq L\|x - y\|.$$

1. For all $p \in \operatorname{conv}\mathcal{P}$, $\left|\frac{p(x)}{p(y)} - 1\right| \leq L\|x - y\|$.
2. For all $p, p_0 \in \operatorname{conv}\mathcal{P}$, $\frac{2p}{p+p_0}$ is $O(L^2)$-Lipschitz wrt $\|\cdot\|$.
3. $\|P - P_n\|_G \stackrel{as}{\to} 0$, where

$$G = \left\{\frac{2p}{p + p_0} : p \in \operatorname{conv}\mathcal{P}\right\}.$$

Non-parametric maximum likelihood: Example

But notice that the dependence on the dimension d is terrible: the rate is exponentially slow in d. The Lipschitz property is a very weak restriction.