Theoretical Statistics. Lecture 16.

Peter Bartlett

1. M-estimators: Consistency of nonparametric maximum likelihood.
2. Asymptotic normality: Delta method.
3. Asymptotic normality of Z-estimators: classical conditions.

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:

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

Non-parametric maximum likelihood

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

Proof:

Because $\hat{p}_n$ maximizes the log likelihood over $\mathcal{P}$, and convexity of $\mathcal{P}$ implies $\tilde{p}_n \in \mathcal{P}$, we have

$$0 \leq \int \log \frac{\hat{p}_n}{\tilde{p}_n} \, dP_n.$$

Now the inequality $\log x \leq x - 1$ implies

$$\begin{aligned} \cdots &\leq \int \left(\frac{\hat{p}_n}{\tilde{p}_n} - 1 \right) \, dP_n \\ &= \int \frac{\hat{p}_n}{\tilde{p}_n} \, d(P_n - P) + \int \frac{\hat{p}_n}{\tilde{p}_n} \, dP - 1. \end{aligned}$$

Non-parametric maximum likelihood

We can write

$$\int \frac{\hat{p}_n}{\tilde{p}_n}\, dP - 1 = \int \frac{2\hat{p}_n}{\hat{p}_n + p_0}\, dP - \int \frac{\hat{p}_n + p_0}{\hat{p}_n + p_0}\, dP$$
$$= -\int \frac{p_0 - \hat{p}_n}{\hat{p}_n + p_0}\, dP.$$

Also,

$$\int \frac{p_0 - \hat{p}_n}{\hat{p}_n + p_0}\, dP = \int \frac{p_0 - \hat{p}_n}{\hat{p}_n + p_0} p_0\, d\mu$$
$$= \frac{1}{2} \int \frac{p_0 - \hat{p}_n}{\hat{p}_n + p_0} (p_0 + \hat{p}_n + p_0 - \hat{p}_n)\, d\mu$$
$$= \frac{1}{2} \int \frac{(p_0 - \hat{p}_n)^2}{\hat{p}_n + p_0}\, d\mu.$$

Non-parametric maximum likelihood

Finally,

$$\left(\hat{p}_n^{1/2} - p_0^{1/2}\right)^2 \left(\hat{p}_n^{1/2} + p_0^{1/2}\right)^2 = (\hat{p}_n - p_0)^2,$$

So

$$\left(\hat{p}_n^{1/2} - p_0^{1/2}\right)^2 = \frac{(\hat{p}_n - p_0)^2}{\left(\hat{p}_n^{1/2} + p_0^{1/2}\right)^2} \leq \frac{(\hat{p}_n - p_0)^2}{\hat{p}_n + p_0}.$$

And this implies

$$h(\hat{p}_n, p_0)^2 \leq \frac{1}{2} \int \frac{(p_0 - \hat{p}_n)^2}{\hat{p}_n + p_0} \, d\mu.$$

Combining gives the result.

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

Consider $\mathcal{P}$ the class of piecewise-polynomial densities (splines) on $[0, 1]$ with $k(n)$ fixed knots (boundaries between the pieces). Since the knots are fixed, it is a convex class. Plus, the ratio class G can be computed in $O(k(n))$ time, so the VC-dimension of the subgraph class is $d = O(k(n)^2)$, and hence

$$\mathbf{E}\|R_n\|_G \leq \sqrt{\frac{d}{n}}.$$

For $k(n) = o(\sqrt{n})$, we have $\|P - P_n\|_G \stackrel{as}{\to} 0$.

(The subgraph class for a class F of real-valued functions is $F_{\leq} = \{(x, y) \mapsto 1[f(x) \geq y] : f \in F\}$. The log covering numbers of F are bounded above by $\log(N(\epsilon, F)) = O(d_{VC}(F_{\leq}) \log(1/\epsilon))$, so computing the entropy integral gives the bound $\|P - P_n\|_F = O(\sqrt{d_{VC}(F_{\leq})/n}$.)

Asymptotic distributions: Delta method

(See vdV3.)

We have an estimator T_n, and we know $T_n \xrightarrow{P} \theta$ and $\sqrt{n}(T_n - \theta) \rightsquigarrow Z$.

Suppose we are interested in the saymptotics of $\phi(T_n)$.

The continuous mapping theorem implies that if ϕ is continuous then $\phi(T_n) \xrightarrow{P} \phi(\theta)$. The delta method gives its asymptotic distribution.

Delta method

Theorem: If $\phi : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable at θ, and $\sqrt{n}(T_n - \theta) \rightsquigarrow T$, then

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T)$$

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) - \phi'_\theta(\sqrt{n}(T_n - \theta)) \xrightarrow{P} 0.$$

Here, ϕ'_θ is the derivative (linear map) satisfying

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o(\|h\|)$$

for $h \to 0$.

Delta method

Proof:

Consider the remainder function

$$R(h) = \phi(\theta + h) - \phi(\theta) - \phi'_\theta(h).$$

Differentiability implies $R(h) = o(\|h\|)$ as $h \to 0$. But $\sqrt{n}(T_n - \theta)$ converges in distribution, so it is uniformly tight and since $\sqrt{n} \to \infty$, $T_n - \theta \xrightarrow{P} 0$. So we can substitute $h = T_n - \theta$ and get

$$R(T_n - \theta) = o_P(\|T_n - \theta\|).$$

But

$$R(T_n - \theta) = \phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta).$$

Delta method

Thus,

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) = \phi'_\theta(\sqrt{n}(T_n - \theta)) + \sqrt{n}o_P(\|T_n - \theta\|).$$

But $\sqrt{n}o_P(\|T_n - \theta\|) = o_P(\sqrt{n}\|T_n - \theta\|) = o_P(1)$.

This shows that

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) - \phi'_\theta(\sqrt{n}(T_n - \theta)) \xrightarrow{P} 0.$$

The continuous mapping theorem and Slutsky's lemma shows that

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T).$$

Asymptotic normality of Z-estimators

Theorem: Consider

$$\Psi_n(\theta) = P_n\psi_\theta, \qquad \Psi(\theta) = P\psi_\theta.$$

Suppose $\hat{\theta}_n \in \mathbb{R}$ is a zero of Ψ_n, $\theta_0 \in \mathbb{R}$ is a zero of Ψ, $\hat{\theta}_n \xrightarrow{P} \theta_0$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ddot{\Psi}_n(\tilde{\theta}_n)}$$

where $\tilde{\theta}_n = \lambda\hat{\theta}_n + (1-\lambda)\theta_0$ for some $0 \leq \lambda \leq 1$.
If $P\psi_{\theta_0}^2$ exists, $P\dot{\psi}_{\theta_0}$ exists and is non-zero, and $\ddot{\Psi}_n(\tilde{\theta}_n) = O_P(1)$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, P\psi_{\theta_0}^2/(P\dot{\psi}_{\theta_0})^2\right).$$