Chapter 3.4: Rates of convergence

Lasse Vuursteen

July 22, 2020

Setup

We continue with a similar setup as last week:

- The aim is estimation of a parameter $\theta_0 \in \Theta$.
- $\hat{\theta}$ maximizes the map

$$\theta \mapsto \mathbb{P}_n m_{n,\theta}.$$

- θ_0 is typically the maximizer of the asymptotic limit of the above map.

This chapter: We wish to know the rate at which $\hat{\theta}$ approaches θ_0.

Setup

Rate of convergence will be expressed in terms of a map $\theta \mapsto d_n(\theta, \theta_{n,0}) \in [0, \infty)$ and a sequence δ_n of nonnegative numbers.

$$d_n(\hat{\theta}, \theta_{n,0}) = O_P^*(\delta_n)$$

So as an example, this could be the typical parametric rate $\delta_n = n^{-1/2}$, d_n corresponding to the Euclidian norm,

$$\sqrt{n}||\hat{\theta} - \theta_{n,0}||_2 = O_P^*(1).$$

Setup

We generalize the previous setting as follows:

- Sequence of *sieves* $\Theta_n \subset \Theta$.
- $\hat{\theta}_n$ takes values in the sieve Θ_n.
- For example, $\hat{\theta}$ maximizes over Θ_n the map

$$\theta \mapsto \mathbb{P}_n m_{n,\theta}.$$

- To achieve consistent estimation, the sieve Θ_n typically grows dense in Θ as $n \to \infty$.

Setup

In the setup of this chapter, $\hat{\theta}_n$ *need not* maximize $\theta \mapsto \mathbb{P}_n m_{n,\theta}$. Instead we need:

- there exists $\theta_n \in \Theta_n$ such that

$$\mathbb{P}_n m_{n,\hat{\theta}_n} \geq \mathbb{P}_n m_{n,\theta_n} - O_P(\delta_n^2)$$

- **NB:** The above does hold if $\hat{\theta}_n$ does maximize the criterion function.
- This θ_n is close to the true parameter θ_0: $d_n(\theta_n, \theta_0) = O(\delta_n)$; which can be thought of as the distance of the sieve to the true parameter.

Setup

Like in the M- and Z-estimation chapters, the theoretical results are formulated in terms of stochatic processes $\mathbb{M}_n$ and (deterministic) M_n indexed by $\Theta_n \cup \{\theta_0\}$.

- $\mathbb{M}_n$ can be thought of as the criterion function $\{\mathbb{P}_n m_{n,\theta} : \theta \in \Theta_n\}$.
- $M_n(\theta) = P m_{n,\theta}$.

Convergence rate theorem

Theorem (3.4.1)

For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set $\Theta_n \cup \{\theta_{n,0}\}$ and let $\theta \mapsto d_n(\theta, \theta_{n,0})$ be an arbitrary map from Θ_n to $[0,\infty)$. Let $\underline{\delta}_n \geq 0$ and suppose that for every n and $\delta > \underline{\delta}_n$,

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta,\theta_{n,0}) \leq \delta} M_n(\theta) - M_n(\theta_{n,0}) \leq -\delta^2,$$

$$E^* \sup_{\theta \in \Theta_n : d_n(\theta,\theta_{n,0}) \leq \delta} \sqrt{n}\left|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})\right| \lesssim \phi_n(\delta),$$

for increasing functions $\phi_n : [\underline{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. Let $\theta_n \in \Theta_n$ and let δ_n satisfy

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2, \quad \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \quad \delta_n \geq \underline{\delta}_n.$$

If the sequence $\hat{\theta}_n$ takes its values in Θ_n and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(\delta_n^2)$, then $d_n(\hat{\theta}_n, \theta_{n,0}) = O_P^*(\delta_n)$.

Convergence rate theorem

Proof: Define shells

$$S_{j,n} := \{\theta \in \Theta_n : \delta_n 2^{j-1} < d_n(\theta, \theta_{n,0}) \leq 2^j \delta_n\}.$$

If $\theta \in S_{n,j}$ and $2^j \delta_n \geq \bar{\delta}_n$,

$$\begin{cases} \sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} M_n(\theta) - M_n(\theta_{n,0}) \leq -\delta^2 \\ M_n(\theta_{n,0}) - M_n(\theta_n) \leq \delta_n^2 \end{cases}$$

imply

$$\begin{aligned} M_n(\theta) - M_n(\theta_n) &= M_n(\theta) - M_n(\theta_{n,0}) + M_n(\theta_{n,0}) - M_n(\theta_n) \\ &\leq -2^{2j}\delta_n^2 + \delta_n^2 \lesssim -2^{2j}\delta_n^2. \end{aligned}$$

Convergence rate theorem

Let

$$G_n = \sqrt{n}(\mathbb{M}_n - M_n).$$

The assumption

$$\left\{ \mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(\delta_n^2) \right.$$

implies

$$G_n(\hat{\theta}_n) - G_n(\theta_n) \geq \sqrt{n}(M_n(\theta_n) - M_n(\hat{\theta}_n)) - O_P(\sqrt{n}\delta_n^2).$$

For $\theta \in S_{n,j}$, we have $M_n(\theta) - M_n(\theta_n) \lesssim -2^{2j}\delta_n^2$; so if $\hat{\theta}_n \in S_{n,j}$

$$G_n(\hat{\theta}_n) - G_n(\theta_n) \geq \sqrt{n}2^{2j}\delta_n^2 - O_P(\sqrt{n}\delta_n^2).$$

So for a given $\varepsilon > 0$ we have for all j large enough

$$P^*\left(\hat{\theta}_n \in S_{n,j}\right) \leq P^*\left(\sup_{\theta \in S_{n,j}} (G_n(\theta) - G_n(\theta_n)) \geq \sqrt{n}2^{2j-1}\delta_n^2\right) + \varepsilon.$$

Convergence rate theorem

We have assumed

$$\begin{cases} E^* \sup_{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta} \sqrt{n} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})| \lesssim \phi_n(\delta) \\ \theta_n \in \Theta_n \end{cases}$$

By the triangle- and Markov inequalities

$$\begin{aligned} & P^* \left(\sup_{\theta \in S_{n,j}} (G_n(\theta) - G_n(\theta_n)) \geq K \sqrt{n} 2^{2j} \delta_n^2 \right) \\ & \leq \frac{\mathbb{E}^* \sup_{\theta \in S_{n,j}} |G_n(\theta) - G_n(\theta_{n,0})| + \mathbb{E}^* |G_n(\theta_{n,0}) - G_n(\theta_n)|}{K \sqrt{n} 2^{2j} \delta_n^2} \\ & \leq \frac{\phi_n(2^j \delta_n) + \phi_n(d(\theta_n, \theta_{n,0}) \vee \underline{\delta}_n)}{K \sqrt{n} 2^{2j} \delta_n^2}. \end{aligned}$$

Convergence rate theorem

We have also assumed that

$$\begin{cases} \text{for } \delta > \underline{\delta}_n, & \sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} M_n(\theta) - M_n(\theta_{n,0}) \leq -\delta^2, \\ \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \end{cases}$$

so if $d_n(\theta_n, \theta_{n,0}) > \underline{\delta}_n$

$$\sup_{\theta \in \Theta_n : d_n^2(\theta_n, \theta_{n,0})/2 < d_n(\theta, \theta_{n,0}) \leq d_n^2(\theta_n, \theta_{n,0})} M_n(\theta) - M_n(\theta_{n,0}) \leq -d_n^2(\theta_n, \theta_{n,0})$$

which when combined with the second assumption above yields

$$d_n^2(\theta_n, \theta_{n,0}) \leq M_n(\theta_{n,0}) - M_n(\theta) \leq \delta_n^2.$$

We obtain

$$\frac{\phi_n(2^j \delta_n) + \phi_n(d(\theta_n, \theta_{n,0}) \vee \underline{\delta}_n)}{\sqrt{n} 2^{2j-1} \delta_n^2} \leq \frac{2\phi_n(2^j \delta_n)}{\sqrt{n} 2^{2j-1} \delta_n^2}$$

Convergence rate theorem

Since ϕ_n is increasing, we now have

$$\phi_n(d(\theta_n, \theta_{n,0}) \vee \underline{\delta}_n) \leq \phi_n(\delta_n).$$

Since $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$, we have $\phi_n(2^j \delta_n) \leq 2^{\alpha j} \phi_n(\delta_n)$. Combining this with $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$,

$$\frac{2\phi_n(2^j \delta_n)}{\sqrt{n} 2^{2j} \delta_n^2} \leq \frac{1}{2^{j(2-\alpha)-2}}.$$

We now have that for $J \in \mathbb{N}$ large enough,

$$P^* \left(d_n(\hat{\theta}, \theta_{n,0}) \geq 2^J \delta_n \right) \leq \sum_{j \geq J} P^* \left(\hat{\theta}_n \in S_{n,j} \right)$$

$$\leq \sum_{j \geq J} \frac{1}{2^{j(2-\alpha)-2}}$$

which is what we wanted to show. ■

Convergence rate theorem

We can apply the previous Theorem to

$$\theta_{n,0} = \arg\max_{\theta \in \Theta} M_n(\theta)$$

and

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta_n} \mathbb{M}_n(\theta).$$

"Squared bias condition":

$$M_n(\theta_{0,n}) - M_n(\theta_n) \leq \delta_n^2$$

Smaller sieves $\Theta_n \subset \Theta$ could increase the square bias.

Convergence rate theorem

Larger sieves $\Theta_n \subset \Theta$ could have implications for the conditions

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} M_n(\theta) - M_n(\theta_{n,0}) \leq -\delta^2,$$

$$E^* \sup_{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta} \sqrt{n} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})| \lesssim \phi_n(\delta),$$

with $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$.

Convergence rate theorem

If a sequence δ_n^2 satisfies the previous two "variance conditions", the proof of the previous theorem shows

$$d_n^2(\hat{\theta}_n, \theta_{n,0}) = O_P^*(\delta_n^2 + M_n(\theta_{n,0}) - M_n(\theta_n)).$$

So the rate of convergence is in the end determined by maximum of the rate in which in the "square bias" and the "variance" decrease.

The continuity modulus

Applying the previous theorem, one needs to derive the continuity modulus of $\sqrt{n}(\mathbb{M}_n - M_n)$ over Θ_n around the truth:

$$E^* \sup_{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta} \sqrt{n} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})| \lesssim \phi_n(\delta).$$

So for this we can use the machinery of chapter 2.14.

The continuity modulus

For example, when $\mathbb{M}_n(\theta) = \mathbb{P}_n m_{n,\theta}$, the classes of functions relevant for the continuity modulus are

$$\mathscr{M}_{n,\delta} := \left\{ m_{n,\theta} - m_{n,\theta_{n,0}} : \theta \in \Theta_n, d_n(\theta, \theta_{n,0}) \leq \delta \right\}.$$

We can bound the continuity modulus in terms of the entropy of these classes.

The continuity modulus

In case we have envelope $M_{n,\delta}$ for the class $\mathscr{M}_{n,\delta}$, we have seen bounds of the form

$$E_P^* ||\mathbb{G}||_{\mathscr{M}_{n,\delta}} \lesssim J(1) \sqrt{P^* M_{n,\delta}^2}$$

with $J(1)$ the uniform or bracketing entropy integral of $\mathscr{M}_{n,\delta}$ (Theorem 2.14.1 or 2.14.15):

$$\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon ||M_{n,\delta}||_{Q,2}, \mathscr{M}_{n,\delta}, L_2(Q))} d\varepsilon.$$

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathscr{M}_{n,\delta}, L_2(P))} d\varepsilon.$$

The continuity modulus

We also have seen bounds of the form

$$E_P^*||\mathbb{G}||_{\mathscr{F}} \lesssim J(\delta,\mathscr{F})\left(1+\frac{J(\delta,\mathscr{F})}{\delta^2\sqrt{n}}\right)$$

(Theorems 2.14.2, 2.14.7-8, 2.14.16-17). Here $\mathscr{F}$ is a class of functions with $f \in \mathscr{F}$ satisfies $||f|| < \delta$ (such as $\mathscr{M}_{n,\delta}$) and

$$J(\delta,\mathscr{F}) = \sup_Q \int_0^\delta \sqrt{\log N(\varepsilon||F||_{Q,2},\mathscr{F},L_2(Q))}d\varepsilon,$$

or

$$J(\delta,\mathscr{F}) = \sup_Q \int_0^\delta \sqrt{\log N_{[]}(\varepsilon,\mathscr{F},L_2(P))}d\varepsilon.$$

The continuity modulus

As stated before $\mathscr{F}$ is a class of functions with $f \in \mathscr{F}$ satisfies $||f|| < \delta$. The "loss function" d_n need not be the norm $||\cdot||$.

So for example we have some class

$$\mathscr{M}_{n,\delta} := \left\{ m_{n,\theta} - m_{n,\theta_{n,0}} : \theta \in \Theta_n, ||m_{n,\theta} - m_{n,\theta_{n,0}}|| \leq \delta \right\}.$$

with $||m_{n,\theta} - m_{n,\theta_{n,0}}|| \leq d_n(\theta, \theta_{n,0})$ which yields:

$$E^* \sup_{\theta : d_n(\theta,\theta_{n,0}) \leq \delta} |\mathbb{G}_n(\theta) - \mathbb{G}_n(\theta_{n,0})| \leq E^* \sup_{\theta : ||m_{n,\theta} - m_{n,\theta_{n,0}}|| \leq \delta} |\mathbb{G}_n(\theta) - \mathbb{G}_n(\theta_{n,0})|$$

We will see this later in the example of maximum likelihood estimators.

Penalized maximum contrast estimators

Penalized maximum contrast estimators maximize a criterion function of the form

$$\theta \mapsto \mathbb{M}_n(\theta) - \hat{\lambda}_n^2 \mathscr{J}_n^2(\theta),$$

$\mathscr{J}_n : \Theta_n \to [0, \infty)$ deterministic, $\hat{\lambda}_n$ a nonnegative random variable defined on the same probability space as $\mathbb{M}_n$.

Penalized maximum contrast estimators

Theorem

For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set $\Theta_n \cup \{\theta_{n,0}\}$ and let $\theta \mapsto d_n(\theta, \theta_{n,0})$ be an arbitrary map from Θ_n to $[0,\infty)$. Let $\underline{\delta}_n \geq 0$, $\lambda_n > 0$ and suppose that for every n and $\delta > \underline{\delta}_n$,

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta,\theta_{n,0}) \leq \delta} M_n(\theta) - M_n(\theta_{n,0}) \leq -\delta^2,$$

$$E^* \sup_{\substack{\theta \in \Theta_n : d_n(\theta,\theta_{n,0}) \leq \delta \\ \mathscr{J}_n(\theta) < \delta/\lambda_n}} \sqrt{n}\left|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})\right| \lesssim \phi_n(\delta),$$

for increasing functions $\phi_n : (\underline{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. Let $\theta_n \in \Theta_n$ and let δ_n satisfy

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2, \quad \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \quad \delta_n \geq \underline{\delta}_n.$$

If $\lambda_n/\hat{\lambda}_n = O_P(1)$ and $\hat{\theta}_n$ maximizes the penalized criterion over Θ_n, then $d_n(\hat{\theta}_n, \theta_{n,0}) = O_P^*(\delta_n + \hat{\lambda}_n \mathscr{J}(\theta_n))$ and $\mathscr{J}_n(\hat{\theta}_n) = O_P^*(\delta_n/\hat{\lambda}_n + \mathscr{J}_n(\theta_n))$.

Penalized maximum contrast estimators

$$d_n(\hat{\theta}_n, \theta_{n,0}) = O_P^*(\delta_n + \hat{\lambda}_n \mathscr{J}(\theta_n))$$

so rate of convergence is $\max\{\delta_n, \hat{\lambda}_n \mathscr{J}(\theta_n))\}$.

- For fast convergence, we would like $\hat{\lambda}_n$ to be small.
- However, $\hat{\lambda}_n$ small implies λ_n small which in turn increases the set over which we take the continuity modulus:

$$E^* \sup_{\substack{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta \\ \mathscr{J}_n(\theta) < \delta/\lambda_n}} \sqrt{n} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})| \lesssim \phi_n(\delta).$$

- The latter yields and increase in δ_n .

An optimal choice of $\hat{\lambda}_n$ balances these effects.

Penalized maximum contrast estimators

Proof (sketch): Define for $M \in \mathbb{N}$, $j \in \mathbb{N}$ the shells

$$S_{n,j} := \{(\theta,\lambda) \in \Theta_n \times [\lambda_n,\infty) \,:\, 2^{2j-2}\delta_n^2 < d_n^2(\theta,\theta_{n,0}) + \lambda^2 \mathscr{J}_n^2(\theta) \leq 2^{2j}\delta_n^2, \\ 2^{2M}\lambda^2 \mathscr{J}_n^2(\theta_n) < d_n^2(\theta,\theta_{n,0}) + \lambda^2 \mathscr{J}_n^2(\theta)\}$$

Combining the displays in the Theorem, it follows that for $(\theta,\lambda) \in S_{n,j}$,

$$M_n(\theta) - M_n(\theta_n) - \lambda^2 \mathscr{J}_n^2(\theta) + \lambda^2 \mathscr{J}_n^2(\theta_n) \lesssim -2^{2j}\delta_n^2.$$

Penalized maximum contrast estimators

Let $G_n := \sqrt{n}(\mathbb{M}_n - M_n)$. By the assumptions on $\hat{\theta}_n$ and $\hat{\lambda}_n$,

$$G_n(\hat{\theta}_n) - G_n(\theta) \geq \sqrt{n}\left[M_n(\theta_n) - M_n(\hat{\theta}_n) - \hat{\lambda}_n^2 \mathscr{J}_n^2(\hat{\theta}_n) + \hat{\lambda}_n^2 \mathscr{J}_n^2(\hat{\theta}_n)\right].$$

If $(\hat{\theta}_n, \hat{\lambda}_n) \in S_{n,j}$, the RHS is greater than $\sqrt{n}2^{2j}\delta_n^2$. So

$$P^*((\hat{\theta}_n, \hat{\lambda}_n) \in S_{n,j}) \leq P^*\left(\sup_{\substack{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta \\ \mathscr{J}_n(\theta) < \delta/\lambda_n}} G_n(\theta) - G_n(\theta_n) \gtrsim \sqrt{n}2^{2j}\delta_n^2\right).$$

From here the proof follows as before: Markov's inequality and the modulus of continuity bound, properties of ϕ_n and summing over $j > M$. (■)

Maximum likelihood estimation of densities

Let $X_1, \ldots, X_n$ an iid sample from a density $p \in \mathscr{P}$ wrt a measure μ on a space $\mathscr{X}$. Consider sieves $\mathscr{P}_n \subset \mathscr{P}$ and suppose $\hat{p}_n$ maximizes $p \mapsto \mathbb{P}_n \log p$ over $\mathscr{P}_n$.

Let P_0 be the distribution corresponding to $p_0 \in \mathscr{P}$ and let $p_n \in \mathscr{P}_n$ be such that $p_0/p_n \leq M$.

For $p_n \in \mathscr{P}_n$, define the criterion function

$$m_{n,p} = \log \frac{p + p_n}{2p_n}.$$

$\mathbb{P}_n m_{n,\cdot}$ plays the role of $\mathbb{M}_n$.

Maximum likelihood estimation of densities

Define squared Hellinger distance to be

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

Define the Bernstein "norm"

$$||f||_{P,B} := \sqrt{2P(e^{|f|} - 1 - |f|)}.$$

Maximum likelihood estimation of densities

Note $m_{n,p_n} = 0$. Since $\hat{p}_n$ maximizes $p \mapsto \mathbb{P}_n \log p$ over $\mathscr{P}_n$ and $p_n \in \mathscr{P}_n$,

$$\mathbb{P}_n \log \frac{\hat{p}_n + p_n}{2p_n} = \mathbb{P}_n \left[\frac{1}{2} \log \hat{p}_n - \frac{1}{2} \log p_n \right] \geq 0.$$

So

$$\mathbb{P}_n \log \frac{\hat{p}_n + p_n}{2p_n} \geq \mathbb{P}_n \log \frac{p_n + p_n}{2p_n} = 0.$$

This means the condition $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(\delta_n^2)$ is satisfied.

Maximum likelihood estimation of densities

As our centering function M_n from before, we take

$$p \mapsto P_0 m_{n,p}.$$

The class for which we bound the continuity modulus is

$$\mathscr{M}_{n,\delta} := \{m_{n,p} - m_{n,p_n} : p \in \mathscr{P}_n \text{ and } h(p, p_n) < \delta\}.$$

Maximum likelihood estimation of densities

Lemma (3.4.6)

Let $p_n \in \mathscr{P}_n$ be such that $p_0/p_n \leq M$. For p such that $h(p,p_n) \geq 32Mh(p_n,p_0)$ it holds that

$$P_0(m_{n,p} - m_{n,p_n}) \lesssim -h^2(p,p_n).$$

Furthermore,

$$E^*_{P_0}||\sqrt{n}(\mathbb{P}_n m_{n,\cdot} - P_0 m_{n,\cdot})||_{\mathscr{M}_{n,\delta}} \lesssim \sqrt{M}\tilde{J}_{[]}(\delta,\mathscr{P}_n,h)\left(1 + \frac{\sqrt{M}\tilde{J}_{[]}(\delta,\mathscr{P}_n,h)}{\delta^2\sqrt{n}}\right)$$

Proof:

- A lot of rewriting and basic inequalities yield the first statement.
- The second follows directly from Theorem 2.14.17.

With this lemma in hand, we can now aim to apply Theorem 3.4.1.

Maximum likelihood estimation of densities

For all $h(p, p_n) \geq 32Mh(p_n, p_0)$,

$$P_0(m_{n,p} - m_{n,p_n}) \lesssim -h^2(p, p_n),$$

which yields that the conditions of the centering function are satisfied for the choices

- $p \mapsto h(p, p_0)$ in place of $\theta \mapsto d_n(\theta, \theta_{n,0}$
- $\delta_n = h(p_n, p_0)$.
- $\underline{\delta}_n = 0$.

Theorem (3.4.1)

For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set $\Theta_n \cup \{\theta_{n,0}\}$ and let $\theta \mapsto d_n(\theta, \theta_{n,0})$ be an arbitrary map from Θ_n to $[0,\infty)$. Let $\underline{\delta}_n \geq 0$ and suppose that for every n and $\delta > \underline{\delta}_n$,

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta,\theta_{n,0}) \leq \delta} M_n(\theta) - M_n(\theta_{n,0}) \leq -\delta^2,$$

$$E^* \sup_{\theta \in \Theta_n : d_n(\theta,\theta_{n,0}) \leq \delta} \sqrt{n}|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})| \lesssim \phi_n(\delta),$$

for increasing functions $\phi_n : [\underline{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. Let $\theta_n \in \Theta_n$ and let δ_n satisfy

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2, \quad \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \quad \delta_n \geq \underline{\delta}_n.$$

If the sequence $\hat{\theta}_n$ takes its values in Θ_n and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(\delta_n^2)$, then $d_n(\hat{\theta}_n, \theta_{n,0}) = O_P^*(\delta_n)$.

Maximum likelihood estimation of densities

$$E^*_{P_0}||\sqrt{n}(\mathbb{P}_n m_{n,\cdot} - P_0 m_{n,\cdot})||_{\mathscr{M}_{n,\delta}} \lesssim \sqrt{M}\tilde{J}_{[]}(\delta, \mathscr{P}_n, h)\left(1 + \frac{\sqrt{M}\tilde{J}_{[]}(\delta, \mathscr{P}_n, h)}{\delta^2\sqrt{n}}\right)$$

means that the condition

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2, \quad \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \quad \delta_n \geq \underline{\delta}_n$$

is satisfied for

$$\phi(\delta) = \tilde{J}_{[]}(\delta, \mathscr{P}_n, h)\left(1 + \frac{\tilde{J}_{[]}(\delta, \mathscr{P}_n, h)}{\delta^2\sqrt{n}}\right)$$

Maximum likelihood estimation of densities

The conditions that are left to be satisfied are

$$\phi(\delta) = \tilde{J}_{[]}(\delta, \mathscr{P}_n, h)\left(1 + \frac{\tilde{J}_{[]}(\delta, \mathscr{P}_n, h)}{\delta^2\sqrt{n}}\right) \leq \sqrt{n}\delta_n^2$$

where $\delta_n^2 = h(p_n, p_0)$. So that means,

$$\tilde{J}_{[]}(\delta_n, \mathscr{P}_n, h) \leq \sqrt{n}\delta_n^2.$$

Maximum likelihood estimation of densities

Example 3.4.10: Monotone densities. Suppose $\mathscr{X} = [0,T]$ and p_0 is known to be nonincreasing.

According to Theorem 2.7.7, the set $\mathscr{F}$ of all nonincreasing functions $f : [0,T] \to [0,1]$ has log bracketing entropy of order $1/\varepsilon$ for the $L_2(\mu)$ norm of a finite measure μ.

If f is nonincreasing, so if $f^{1/2}$, so the Hellinger distance "qualifies" as the L_2 norm on the roots of the nonnegative functions.

$$\tilde{J}_{[]}(\delta_n, \mathscr{P}_n, h) = \sup_Q \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathscr{F}, L_2(Q))} d\varepsilon = \delta^{1/2}$$

Maximum likelihood estimation of densities

We have

$$\delta_n^{1/2} = \tilde{J}_{[]}(\delta_n, \mathscr{P}_n, h) \lesssim \sqrt{n}\delta_n^2$$

which yields

$$\delta_n \gtrsim n^{-1/3}.$$

Similarly, the chapter discusses convex densities (Example 3.4.11) and densities consisting of Gaussian mixtures (Example 3.4.12).

Least squares regression with fixed design

Given fixed design $x_1, \ldots, x_n$ in a set $\mathscr{X}$ and a map $\theta_0 : \mathscr{X} \to \mathbb{R}$ let

$$Y_i = \theta_0(x_i) + e_i$$

where $e_1, \ldots, e_n$ iid mean zero random variables.

The least squares estimator $\hat{\theta}_n$ minimizes the criterion

$$\theta \mapsto \mathbb{P}_n(Y - \theta)^2$$

If we insert $Y_i = \theta_0(x_i) + e_i$, we see that $\hat{\theta}_n$ maximizes

$$\theta \mapsto 2\mathbb{P}_n(\theta - \theta_0)e - \mathbb{P}_n(\theta - \theta_0)^2.$$

We do not observe these two terms, but we can apply Theorem 3.4.1 to this criterion function to derive the convergence rate.

Least squares regression with fixed design

Since $\mathbb{E}e_i = 0$, the criterion

$$2\mathbb{P}_n(\theta - \theta_0)e - \mathbb{P}_n(\theta - \theta_0)^2$$

is "centered by"

$$M_n(\theta) := \mathbb{P}_n(\theta - \theta_0)^2.$$

The continuity modulus of interest then is then

$$E^* \sup_{\mathbb{P}_n(\theta-\theta_0)^2 < \delta^2, \theta \in \Theta_n} \sqrt{n}|\mathbb{P}_n(\theta - \theta_0)e|.$$

Least squares regression with fixed design

If e_i is sub-Gaussian, note that

$$\theta \mapsto \sqrt{n}\mathbb{P}_n(\theta - \theta_0)e$$

is itself a sub-Gaussian process for the $L_2(\mathbb{P}_n)$ metric. We may apply

Corollary (2.2.8)

Let $\{X_t : t \in T\}$ separable sub-Gaussian process for metric d. Then we have for all $t_0 \in T$

$$\mathbb{E}\sup_t |X_t| \lesssim \mathbb{E}|X_{t_0}| + \int_0^\infty \sqrt{\log N(\varepsilon, T, d)}d\varepsilon$$

Least squares regression with fixed design

This yields

$$E^* \sup_{\mathbb{P}_n(\theta-\theta_0)^2<\delta^2,\theta\in\Theta_n} \sqrt{n}|\mathbb{P}_n(\theta-\theta_0)e| \lesssim \phi_n(\delta).$$

for

$$\phi_n(\delta) = \int_0^\delta \sqrt{\log N(\varepsilon,\Theta_n,L_2(\mathbb{P}_n))}d\varepsilon.$$

So for "sufficiently small" sieves we can obtain good rates in terms of this entropy integral.

For errors with sub-Gaussian tails where the regression function is known to belong to $C_1^\alpha[K]$ for $K \subset \mathbb{R}^d$ a compact set, we achieve the rate

$$\delta_n = n^{\frac{\alpha}{2\alpha+d}}$$

in the $L_2(\mathbb{P}_n)$ semimetric (see Example 3.4.14).

Least squares regression with fixed design

Another example of a sieving strategy is to take a finite dimensional space that grows towards the full space of interest.

$$\Theta_n = \text{span}(\psi_1, \ldots, \psi_{N_n}).$$

A direct computation yields

$$E^* \sup_{\mathbb{P}_n(\theta-\theta_0)^2 < \delta^2, \theta \in \Theta_n} \sqrt{n}|\mathbb{P}_n(\theta - \theta_0)e| \leq \delta_n^2 N_n E e_i^2.$$

So errors with finite second moments and small enough N_n leads to a small modulus of continuity.

To keep N_n small, one can choose a basis $\{\psi_1, \ldots\}$ that approximates the truth well (e.g. not needing too many basis elements to do so) (see Examples 3.4.15 and Example 3.4.16).

More examples in the chapter:

- 3.4.7 - Regression with random design
- 3.4.8 - Penalized least-squares Regression
- 3.4.9 - Least absolute deviation regression
- 3.4.10 - Lasso
- 3.4.11 - Classification

Thank you for listening!