

Notes on selected topics in Statistics

Krishnakumar Balasubramanian

WARNING: This note is primarily written by the author for self-understanding. It may undergo frequent revisions. Please notify me if you find any typos.

Contents

1	Random variables, distribution and density functions	2
2	Stochastic order notation	2
3	Metrics	3
4	CLT without a variance	4
5	Evaluating density estimates	4
6	Asymptotic statistics	4
7	Likelihood ratio test	8
8	Minmax lower bounds for density estimation	10
9	Structural assumptions made for high dimensional estimation	12
10	Conditional expectations and martingales	12
11	Entropy of Metric Spaces	13
12	Rate of convergence of MLE via Empirical process	14
13	Influence functions and efficient estimation	16

1 Random variables, distribution and density functions

Definition 1 (Probability spaces). A probability space is a space (Ω, \mathcal{F}, P) where Ω is the set of possible outcomes, \mathcal{F} is the set of all events, $P : \mathcal{F} \rightarrow [0, 1]$ is a function that assigns probabilities to the events. We operate under the assumption \mathcal{F} is a σ -field, i.e., (i) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ and (ii) if $A_i \in \mathcal{F}$ is a countable sequence of sets then $\cup_i A_i \in \mathcal{F}$

Borel set is any set in a topological space that can be formed from open sets (or, equivalently, from closed sets) through the operations of countable union, countable intersection, and relative complement.

Note that for any measure μ , $\mu(\cup_i A_i) = \sum_i \mu(A_i)$ is satisfied. If $\mu(\Omega) = 1$, then that measure is called as probability measure. Consider the real line \mathbb{R} with borel sets \mathcal{R} and a non-decreasing and right continuous function F . Associated with every such function F is a unique measure μ on $(\mathbb{R}, \mathcal{R})$ with $\mu((a, b]) = F(b) - F(a)$. If $F(x) = x$, then the resulting measure is Lebesgue measure. The probability spaces become more interesting when we define random variables on them.

Definition 2 (Random variable). A real valued function $X : \Omega \rightarrow \mathbb{R}$ is called a random variable if for every Borel set $B \subset \mathbb{R}$, we have $X^{-1}(B) = \{w : X(w) \in B\} \in \mathcal{F}$.

Definition 3 (Distribution function of random variable). If X is a random variable, it induces a probability measure on \mathbb{R} , called its distribution by setting $\mathbb{P}(A) = P(X \in A) = P(X^{-1}(A))$ for borel sets A , i.e., we pull back $A \in \mathbb{R}$ to $X^{-1}(A) \in \mathcal{F}$ and then take the P of that set. The distribution of a random variable X is usually described by giving its distribution function, $F(x) = \mathbb{P}(X \leq x)$.

As shorthand, one writes the probability $\mathbb{P}(A) = P(X \in A)$. In fact, its common not to write the induced measure at all, and just write $P(X \in A)$. A distribution function satisfies the following:

1. F is non-decreasing
2. $\lim_{x \rightarrow \infty} F(x) = 1$ and $\lim_{x \rightarrow -\infty} F(x) = 0$
3. F is right continuous with left limits.
4. $\mathbb{P}(X = x) = F(x) - F(x^-)$

When $F(x)$ could be written as $F(x) = \int_{-\infty}^x f(y) dy$, we say that X has a density function, $f(x)$, where f is a non-negative Lebesgue measurable function.

2 Stochastic order notation

Definition 4 (Small o). Let $\{X_n\}$ be a sequence of random variables. Then $X_n = o_p(1) \implies X_n \rightarrow_p 0$.

Definition 5 (Big O). Let $\{X_n\}$ be a sequence of random variables. Then $X_n = O_p(1) \implies X_n$ is stochastically bounded i.e.,

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} P(|X_n| > M) = 0.$$

Properties:

- $o_p(1) + p_p(1) = o_p(1)$
- $o_p(1) + O_p(1) = O_p(1)$
- $O_P(1)o_p(1) = o_p(1)$
- $o_p(O_p(1)) = o_p(1)$
- $(1 + o_p(1))^{-1} = O_p(1)$
- $\{E|X_n|^k\}$ is bounded for some $k \geq 1 \implies X_n = O_p(1)$
- $X_n = O_p(n^{-\beta}) \implies X_n = o_p(\alpha_n)$ for any sequence $\{\alpha_n\}$ with $n^\beta \alpha_n \rightarrow \infty$ as $n \rightarrow \infty$.

- $X_n \rightarrow_D X \implies X_n = O_p(1)$
- $X - n \rightarrow_D X$ and $|X_n - Y_n| \rightarrow_p 0 \implies Y_n \rightarrow_D X$

Theorem 2.1 (Polyas Theorem). *If $F_n \rightarrow_D F$, and F is a continuous CDF, then $\|F_n - F\|_\infty \rightarrow 0$ as $n \rightarrow \infty$.*

Theorem 2.2 (Slutsky's Theren). • $X_n \rightarrow_D X$ and $Y_n \rightarrow_p c$, then $X_n \cdot Y_n \rightarrow_D cX$.

- $X_n \rightarrow_D X$ and $Y_n \rightarrow_p c \neq 0$, then $\frac{X_n}{Y_n} \rightarrow_D \frac{X}{c}$.
- $X_n \rightarrow_D X$ and $Y_n \rightarrow_p c$, then $X_n + Y_n \rightarrow_D X + c$

Theorem 2.3 (Borel-Cantelli). *Let $\{A_n\}$ be a sequence of events on a probability space. If $\sum_{n=1}^\infty P(A_i) < \infty$, then $P(\text{infinitely many } A_n \text{ occur}) = 0$. Additionally, if $\{A_n\}$ are pairwise independent and $\sum_{n=1}^\infty P(A_i) = \infty$, then $P(\text{infinitely many } A_n \text{ occur}) = 1$.*

Below we give an example of how the above Borel-Cantelli lemma could be used for proving almost-sure convergence.

- Consider $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$ and suppose \bar{X}_n is the mean, For a fixed $\epsilon > 0$, by Markov's inequality, $P(|\bar{X}_n| > \epsilon) \leq \frac{E(\bar{X}_n^4)}{\epsilon^4} = \frac{3}{\epsilon^4 n^2}$. Since $\sum_{i=1}^\infty \frac{1}{n^2} < \infty$, $P(|\bar{X}_n| > \epsilon \text{ infinitely often}) = 0$ and hence $\bar{X}_n \rightarrow a.s0$

Theorem 2.4. *Suppose $E(|X_n - c|) \rightarrow 0$ for some $-\infty < c < \infty$. By Markov's inequality, $P(|X_n - c| > \epsilon) \leq \frac{E(|X_n - c|)}{\epsilon} \rightarrow 0$ as $n \rightarrow \infty$. Thus $X_n \rightarrow_p c$.*

3 Metrics

- Metric for convergence in prob: $d_E(X, Y) = E\left(\frac{|X - Y|}{1 + |X - Y|}\right)$.
- Kolmogorov Metric: $d_K(F, G) = \sup_x |F(x) - G(x)|$
- Levy Metric: $d_L(F, G) = \inf\{\epsilon > 0 : F(x - \epsilon) - \epsilon \leq G(x) \leq F(x + \epsilon) + \epsilon \quad \forall x\}$
- Total Variation metric: Let P and Q be absolutely continuous with respect to some measure μ , then $d_{TV}(P, Q) = \frac{1}{2} \int |f(x) - g(x)| d\mu$, where f and g are densities of P and Q w.r.t μ .
- KL Divergence: $K(P, Q) = - \int (\log \frac{q}{p}) p d\mu$.
- Hellinger: $H^2(P, Q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu$
- $H(P, Q) \leq \sqrt{K(P, Q)}$
- $H(P, Q) \geq d_{TV}(P, Q)$
- $\frac{H(P, Q)}{\sqrt{2}} \leq \sqrt{d_{TV}(P, Q)}$ and hence $d_{TV}(P, Q) \rightarrow 0 \iff H(P, Q) \rightarrow 0$.
- Convergence in Kullback-Leibler distance implies convergence in total variation and hence convergence in law.

Definition 6. Below we see some properties of different types of convergences in terms of the metrics defined above.

- $X_n \rightarrow_p X \iff d_E(X_n, X) \rightarrow 0$.
- $X_n \rightarrow_D X \iff d_L(F_n, F) \rightarrow 0$ where $X_n \sim F_n, X \sim F$.
- $d_K(F_n, F) \rightarrow 0 \implies X_n \rightarrow_D X$. Reverse is true only under additional conditions.
- $d_{TV}(P_n, P) \rightarrow 0 \implies X_n \rightarrow_D X$. Converse not necessarily true.

The Kullback-Leibler distance is very popular in statistics. Specifically, it is frequently used in problems of model selection, testing for goodness of fit, Bayesian modeling and Bayesian asymptotics, and in certain estimation methods known as minimum distance estimation. The Kolmogorov distance is one of the easier ones computationally and has been used in many problems, too, and notably so in the literature on robustness and Bayesian robustness. The Hellinger distance is a popular one in problems of density estimation and in time series problems. The Levy metric is technically hard to work with but metrizes weak convergence, a very useful property. It, too, has been used in the robustness literature, but it is more common in probability theory. Convergence in total variation is extremely strong, and many statisticians seem to consider it unimportant. But it has a direct connection to L_1 distance, which is intuitive. It has a transformation invariance property and, when it holds, convergence in total variation is extremely comforting.

4 CLT without a variance

This is a version of CLT which does not require finiteness of variance. First, we start with a definition.

Definition 7. A function $L : \mathcal{R} \rightarrow \mathcal{R}$ is called slowly varying at ∞ if, for every $t > 0$, we have $\lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = 1$.

For example, functions $\log x$, $\frac{x}{1+x}$ and any function with a finite limit as $x \rightarrow \infty$ satisfy the above definition. But $\exp x$ is not slowly varying.

Theorem 4.1. let $X_1, X_2 \dots \stackrel{\text{iid}}{\sim} F$, where F is a CDF on \mathcal{R} . Let $v(x) = \int_{[-x,x]} y^2 dF(y)$. Then, there exist constants $\{a_n\}$, $\{b_n\}$ such that, $\frac{\sum_i X_i - a_n}{b_n} \rightarrow_D N(0, 1)$, if and only if $v(x)$ is slowly varying at ∞ .

If F has finite second moment, then $v(x)$ is slowly varying at ∞ . Now, lets look at an example. Suppose $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} f(x) = \frac{c}{(2+x^2)^{3/2}}$, a t -distribution with 2 degrees of freedom that has finite mean but not a finite variance. With some algebra, one can see that the distribution has a $v(x)$ which is slowly varying at ∞ . Hence it follows that for i.i.d samples from t -distribution with 2 degrees of freedom, the partial sums converge in distribution to standard normal distribution with $a_n = 0, b_n = \sqrt{n \log n}$.

5 Evaluating density estimates

There are three basic ways to quantify 'good' density estimates.

- $\int (f(x) - \hat{f}(x))^2 dx$ should be small.
- $\int |f(x) - \hat{f}(x)| dx$ should be small. We also have that, $\frac{1}{2} \int |f(x) - \hat{f}(x)| dx$ is the maximum error in our estimate of probability of any set.
- $\int f(x) \log \frac{f(x)}{\hat{f}(x)} dx$ should be small.

Lets expand the third quantity.

$$D(f||\hat{f}) = - \int f(x) \log \hat{f}(x) dx + \int f(x) \log f(x) dx$$

The second term does not involve the density estimate. Hence, approximating the integral in the first term with a summation, $\int f(x) \log \hat{f}(x) \approx \frac{1}{n} \sum_i \log f(\hat{x}_i)$, which is the log-likelihood. If we plug-in the kde, we have

$$\begin{aligned} &= \frac{1}{n} \sum_i \log \left(\frac{1}{nh} \sum_j K\left(\frac{x_j - x_i}{h}\right) \right) \\ &\approx - \log nh + \log K(0) \end{aligned}$$

which we got by taking h to be very small and hence noting that $k\left(\frac{x_j - x_i}{h}\right) \approx 0$ unless $x_i = x_j$. The expression becomes $+\infty$ when $h \rightarrow 0$ which is the maximum of the likelihood. Hence we note that Kde is infact a maximum likelihood estimator as $h \rightarrow 0$. In fact, the limit is to say that $\hat{f}(x) = \frac{1}{n} \sum_i \delta(x - x_i)$ is the form of KDE in that case, which is what we would get if we took the empirical CDF as such.

6 Asymptotic statistics

First we define what is meant by rate of convergence.

Definition 8 (Rate of convergence). If $\hat{\theta}_n$ is an estimator of θ^* , the estimator is consistent with convergence rate δ_n if

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} P(\hat{\theta}_n - \theta^* \geq M\delta_n) = 0$$

in which case, we write $\theta_n - \theta^* = O_p(\delta_n)$.

Equivalently, if $\delta_n^{-1}(\theta_n - \theta^*)$ converge in distribution to some law, by properties of convergence in distribution, we have $\theta_n - \theta^* = O_p(\delta_n)$.

Definition 9 (Quadratic mean differentiability). The family $\{P_\theta, \theta \in \Theta\}$ is quadratic mean differentiable (q.m.d.) at θ_0 , if there exist a vector of real valued function $\eta(\cdot, \theta_0) = [\eta_1(\cdot, \theta_0) \cdots \eta_n(\cdot, \theta_0)]^\top$ such that

$$\int_{\mathcal{X}} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \langle \eta(x, \theta_0), h \rangle \right]^2 d\mu(x) = o(|h|^2)$$

as $|h| \rightarrow 0$

Before some facts about q.m.d., its insightful to compare it with derivatives one studied in calculus. For a function $f(a)$, a derivative must satisfy the property that $\lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - f'(a)h}{h} = 0$ and hence $f(a+h) \approx f(a) + f'(a)h$. In the case of q.m.d., first consider the integrand without the square. For each fixed x , $\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \langle \eta(x, \theta_0), h \rangle = o(|h|)$, which is a much stronger condition. Hence, we sort of relax the requirement to the condition that average with respect to a measure μ is $o(|h|^2)$. Let $L^2(\mu)$ denote space of square integrable functions, i.e., $\int g^2(x) d\mu(x) < \infty$. The convenience of working with square root densities is due to the fact that $\sqrt{p_\theta(x)} \in L^2(\mu)$. If the q.m.d holds good for all θ_0 , then we say the family itself is q.m.d.

Theorem 6.1. Assume that $\{P_\theta, \theta \in \Theta\}$ is quadratic mean differentiable (q.m.d.) at θ_0 . Let $h \in \mathcal{R}^K$. Then

- $\int \sqrt{p_\theta(x)} \langle \eta(x, \theta_0), h \rangle d\mu(x) = 0$ i.e., $\langle \frac{\eta(x, \theta_0)}{\sqrt{p_\theta(x)}}, h \rangle$ is a r.v. with mean 0, under P_θ .
- The components of $\eta(\cdot, \theta_0)$ are in $L^2(\mu)$.
- The finite dimensional set of vectors $\eta(\cdot, \theta_0) \in L^2(\mu)$ is orthogonal to $\sqrt{p_\theta(x)}$.
- For a q.m.d. family with derivative $\eta(\cdot, \theta_0)$, fisher information matrix $I(\theta)_{(i,j)} = 4 \int \eta_i(x, \theta) \eta_j(x, \theta) d\mu(x)$. The existence of FIM follows from step 2.
- For any $h \in \mathcal{R}^k$, $\int |\langle h, \eta(x, \theta_0) \rangle|^2 d\mu(x) = \frac{1}{4} \langle h, I(\theta_0)h \rangle$.
- Sufficient condition for q.m.d.: Point wise differentiability for $p_\theta(x)$ w.r.t. θ , i.e., $\eta_i(x, \theta) = \frac{\partial p_\theta(x) / \partial \theta_i}{2\sqrt{p_\theta(x)}}$

Definition 10. Let P_n, Q_n be probability distributions on $(\mathcal{X}_n, \mathcal{F}_n)$. The sequence Q_n is contiguous to the sequence P_n if $P_n(E_n) \rightarrow 0 \implies Q_n(E_n) \rightarrow 0$ for every sequence E_n with $E_n \in \mathcal{F}_n$.

Useful way to check if Q_n is contiguous w.r.t. P_n : Suppose p_n, q_n are densities of P_n, Q_n w.r.t μ . For $x \in \mathcal{X}$, define

$$L_n(x) = \begin{cases} \frac{q_n}{p_n} & \text{if } p_n(x) > 0 \\ \infty & \text{if } p_n(x) = 0 < q_n(x) \\ 1 & \text{if } p_n(x) = 0 = q_n(x) \end{cases}$$

Note that, L_n is an extended random variable. Observe that,

$$\mathbb{E}_{P_n}(L_n) = Q_n x : p_n(x) > 0 = 1 - Q_n x : p_n(x) = 0 \leq 1$$

with equality if and only if Q_n is absolutely continuous w.r.t to P_n .

For notational convenience, denote $\dot{\ell}_\theta = \frac{2\eta(\cdot, \theta)}{\sqrt{p_\theta}}$. We characterize and define LAN in terms of q.m.d. as it is the weakest condition needed for our purposes. The advantage is that requiring only 1st order derivatives, we have 2nd order Taylor series expansion.

Theorem 6.2 (Local asymptotic normality). Suppose Θ be an open subset of \mathcal{R}^k and that the model $\{P_\theta : \theta \in \Theta\}$ is q.m.d at θ . As a consequence we have $\mathbb{E}_\theta \dot{\ell}_\theta = 0$ and FIM $I_\theta = \mathbb{E}_\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top$ exist. Furthermore, for every converging sequence $h_n \rightarrow h$, as $n \rightarrow \infty$

$$\log \Pi_{i=1}^n \frac{p_{\theta+h_n/\sqrt{n}}(X_i)}{p_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^\top \dot{\ell}_\theta(X_i) - \frac{1}{2} h^\top I_\theta h + o_{p_\theta}(1).$$

Proof sketch. Let consider the case when $h_n = h$ for simplicity. We have

$$\log \prod_{i=1}^n \frac{p_{\theta+h_n/\sqrt{n}}(X_i)}{p_{\theta}(X_i)} = 2 \log \prod_{i=1}^n \sqrt{\frac{p_{\theta+h_n/\sqrt{n}}(X_i)}{p_{\theta}(X_i)}} = 2 \sum_{i=1}^n \log \left(1 + \sqrt{\frac{p_{\theta+h_n/\sqrt{n}}(X_i)}{p_{\theta}(X_i)}} - 1 \right)$$

Fix θ and h and denote $p = p_{\theta}$, $p_n = p_{\theta+h/\sqrt{n}}$ and let $W_{n_j} = 2 \left(\sqrt{\frac{p_n(X_i)}{p(X_i)}} - 1 \right)$ in the above expression to get the simplification $2 \sum_{i=1}^n \log \left(1 + \frac{1}{2} W_{n_j} \right)$. Note that the Taylor series expansion (up to second order) for $\log(1+x) = x - \frac{x^2}{2}$. Hence we get

$$\begin{aligned} 2 \log \left(1 + \frac{1}{2} W_{n_j} \right) &= W_{n_j} - \frac{W_{n_j}^2}{4} + W_{n_j}^2 \rho(W_{n_j}) \\ \log \prod_{i=1}^n \frac{p_{\theta+h_n/\sqrt{n}}(X_i)}{p_{\theta}(X_i)} &= \sum_{i=1}^n W_{n_j} - \sum_{i=1}^n \frac{W_{n_j}^2}{4} + \sum_{i=1}^n W_{n_j}^2 \rho(W_{n_j}) \end{aligned}$$

Note that the first term converges to $\frac{1}{\sqrt{n}} \sum_{i=1}^n h^{\top} \dot{\ell}_{\theta}(X_i) - \frac{1}{4} h^{\top} I_{\theta} h + o_p(1)$, second term to $\frac{1}{4} h^{\top} I_{\theta} h + o_p(1)$ and the third term to $o_p(1)$ as $n \rightarrow \infty$. \square

The definition of LAN is not only restricted to standard i.i.d. case of experiments, but is more general and could be applied to non-i.i.d cases such as time series and random fields as well.

Definition 11. The sequence of statistical model $(P_{n,\theta} : \theta \in \Theta)$ is LAN at θ if there exists matrices r_n and I_{θ} and random vectors $\delta_{n,\theta}$ such that $\delta_{n,\theta} \rightarrow_D N(0, I_{\theta})$ and for every converging sequence $h_n \rightarrow h$

$$\log \frac{dP_{n,\theta+h_n/r_n}}{dP_{n,\theta}}(X_i) = h^{\top} \delta_{n,\theta} - \frac{1}{2} h^{\top} I_{\theta} h + o_{p_{n,\theta}}(1).$$

Next we derive limiting distribution for difference between truth and estimate. By Hodges counter example, one knows that, its impossible to build non-trivial lower bounds on the limiting distribution of estimators uniformly in θ . One can always improve on any given estimator to perform better for selected parameters. So we consider a locally uniform type bound. Consider parameters $\theta + h/\sqrt{n}$ for fixed θ and h ranging over \mathcal{R}^k and suppose for limiting distribution, $L_{\theta,h}$,

$$\sqrt{n} \left(T_n - \phi \left(\theta + \frac{h}{\sqrt{n}} \right) \right) \rightarrow_{D, (\theta+h/\sqrt{n})} L_{\theta,h} \quad \forall h.$$

Then T_n is a good estimate for $\phi(\theta)$ if $L_{h,\theta}$ are maximally concentrated at zero. If P_{θ} depends smoothly on parameter, then

$$(P_{\theta+h/\sqrt{n}}^n : h \in \mathcal{R}^k) \rightarrow_D (N(h, I_{\theta}^{-1}) : h \in \mathcal{R}^k)$$

The proof of the above statement requires contiguity arguments.

Theorem 6.3. Assume experiment P_{θ} is qmd with non-singular I_{θ} . Let ϕ be differentiable at θ . Let T_n be estimators in the experiment $(P_{\theta+h/\sqrt{n}}^n : h \in \mathcal{R}^k)$. Then there exists a randomized statistic T in the experiment $(N(h, I_{\theta}^{-1}) : h \in \mathcal{R}^k)$ such that $T - \dot{\phi}_{\theta} h$ has limiting distribution $L_{\theta,h}$ for every h .

Now consider normal means problem of estimating Ah based on observing single observation X from $N(h, \Sigma)$, Σ known and non-singular. AX is MVU estimator. A randomized estimator T is called equivariant-in-law for estimating Ah if the distribution of $T - Ah$ under h does not depend on h . The law of $AX - Ah$ is $N(0, A\Sigma A^{\top})$ and hence AX is equivariant estimator.

Definition 12. The null distribution of any randomized equivariance in law estimator of Ah could be decomposed as $L = N(0, A\Sigma A^{\top}) * M$, where M is some random measure. Only for AX , $M = 0$. Note that convolving a measure with another measure decreases concentration. More precisely we have the following lemma.

Definition 13 (Anderson's Lemma). For any bowl shaped loss function l on \mathcal{R}^k , every probability measure M on \mathcal{R}^k and every covariance matrix Σ , we have

$$\int l dN(0, \Sigma) \leq \int l d[N(0, \Sigma) * M]$$

A sequence T_n is called regular at θ for estimating $\phi(\theta)$ if for every h ,

$$\sqrt{n} \left(T_n - \phi\left(\theta + \frac{h}{\sqrt{n}}\right) \right) \rightarrow_D (\theta + h/\sqrt{n})L_\theta.$$

A regular estimate attains its limiting distribution in a locally uniform manner. Every estimator sequence is matched by an estimator T in the limit experiment $N(h, I_\theta^{-1})$ and for regular estimate sequence we have $T - \dot{\phi}_\theta h \rightarrow_D L_\theta$ for every h . The best regular estimator sequences is the T_n that corresponds to best equivariant-in-law estimator T for $\dot{\phi}_\theta h$ in the limit experiment, which is $\dot{\phi}_\theta X$ and the best possible limiting distribution is $N(0, \dot{\phi}_\theta I_\theta^{-1} \dot{\phi}_\theta^\top)$

Theorem 6.4 (Convolution). *Assume experiment P_θ is qmd with non-singular I_θ . Let ϕ be differentiable at θ . Let ϕ be differentiable at θ . Let T_n be a regular estimator sequence in the experiment $(P_\theta^n : \theta \in \Theta)$ with limit distribution L_θ . Then, there exists a probability measure M_θ such that*

$$L_\theta = N(0, \dot{\phi}_\theta I_\theta^{-1} \dot{\phi}_\theta^\top) * M_\theta$$

In particular, if L_θ has covariance matrix Σ_θ , then the matrix $\Sigma_\theta - \dot{\phi}_\theta I_\theta^{-1} \dot{\phi}_\theta^\top$ is non-negative definite.

There also exists a version of the above convolution theorem, which says in the above case, the Hodge's kind of improvement could be done only on a parameter set of measure zero. Now, we look at another way (i.e., minimax) of justifying normal limiting distribution. First we will consider the normal means case. According to minimax criterion, the best estimator relative to any given loss function minimized the following maximum risk (over all estimators T)

$$\sup_h E_h l(T - Ah).$$

Proposition 1. *For any bowl shaped loss function, the maximum risk of any randomized estimator T of Ah is bounded below by $E_0 l(AX)$. Consequently, AX is minimax estimator of Ah .*

Under certain additional conditions, we also have AX to be the only minimax estimator.

Now, in the general case, we have the following theorem.

Theorem 6.5 (Local Asymptotic Minimax Theorem). *Let the experiment $(P_\theta : \theta \in \Theta)$ be qmd at θ with non-singular FIM. Let ϕ be differentiable at θ . Let T_n be an estimator sequence in the experiments $(P_\theta^n : \theta \in \mathcal{R}^k)$. Then for any bowl shaped loss function l ,*

$$\sup_I \liminf_{n \rightarrow \infty} \sup_{h \in I} E_{\theta+h/\sqrt{n}} l \left(\sqrt{n} \left(T_n - \phi\left(\theta + \frac{h}{\sqrt{n}}\right) \right) \right) \geq \int l dN(0, \dot{\phi}_\theta I_\theta^{-1} \dot{\phi}_\theta^\top)$$

The integral on the right corresponds to $E_0 l(AX)$, which we saw in the previous proposition and first supremum is taken over all finite subsets of \mathcal{R}^k . This kind of optimality is accepted by most statisticians (even though there exists other estimators, for special cases, which is optimal in some other sense) for the following two reasons.

- First improvement can be made only on a null set of parameters (by almost everywhere convolution theorem).
- Improvements could be done only for some loss functions and improvements wrt one loss function necessarily means worse performance for other loss functions.

Hence, as a trade off between generality and specificity, most statisticians accepts convolution theorem as basis for optimality. Under that assumption, we will see how to achieve the bound set by the theorem or how to build efficient estimators. We will see that an estimator will have the property, only if estimator is asymptotically linear in the score function.

Theorem 6.6. *Let the experiment $(P_\theta : \theta \in \Theta)$ be qmd at θ with non-singular FIM. Let ϕ be differentiable at θ . Let T_n be an estimator sequence in the experiments $(P_\theta^n : \theta \in \mathcal{R}^k)$ such that*

$$\sqrt{n}(T_n - \phi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\phi}_\theta I_\theta^{-1} \dot{\phi}_\theta(X_i) + o_{p_\theta}(1).$$

Then T_n is the best regular estimator for $\phi(\theta)$ at θ . Conversely, every best estimator sequence satisfy this expansion.

This kind of analysis is feasible for parametric models. Similar results could be obtained using large deviations theory for parametric models, which will be useful for non-parametric models too.

7 Likelihood ratio test

Let $\{X_1, \dots, X_n\} \stackrel{\text{iid}}{\sim} P_\theta$, $\theta \in \Theta \subset \mathcal{R}^d$, and open. The likelihood function $L_n(\theta) = \prod_{i=1}^n P_\theta(X_i)$ is a stochastic process parametrized by θ . The maximum likelihood estimator is defined as $\hat{\theta}_n = \arg \max_{\theta \in \Theta} L_n(\theta)$. Under some regularity conditions, on P_θ , we have (i) $\hat{\theta}_n \rightarrow \theta_0$ as $n \rightarrow \infty$ almost surely, (ii) $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_D N(0, I^{-1}(\theta))$, where $I(\theta)$ is the fisher information matrix.

Now consider a testing problem. First let $\Theta = \Theta_0 \cup \Theta_1$ and null hypothesis $H_0 \stackrel{\text{def}}{=} \theta \in \Theta_0 = L \subset \mathcal{R}^d$, where L is a linear subspace, versus $H_a \stackrel{\text{def}}{=} \theta \in \Theta_1$. Consider the statistic (motivated from Neyman-Pearson lemma), defined by

$$\tilde{\Lambda}_n = \log \frac{\sup_{\theta \in \Theta_1} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)}$$

Let $\hat{\theta}_n^0 = \arg \max_{\theta \in \Theta_0} L_n(\theta)$ be the MLE under null hypothesis and $\hat{\theta}_n^a = \arg \max_{\theta \in \Theta_1} L_n(\theta)$, the MLE under alternate hypothesis. Since we do not know the actual values in composite hypothesis case, we use the plug-in estimator calculated by MLE. Hence, we have

$$\tilde{\Lambda}_n = \log \frac{L_n(\hat{\theta}_n^a)}{L_n(\hat{\theta}_n^0)}$$

But, the MLE might be hard to calculate under alternate hypothesis. So, we modify the above statistic by making the supremum in the numerator over the entire space.

$$\begin{aligned} \Lambda_n &\stackrel{\text{def}}{=} 2 \left(\log \frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)} \right) \\ &= 2 \log \frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_n^0)} \end{aligned}$$

We claim that the above statistic converges in distribution to a chi-square distribution with degrees of freedom $d - \dim(L)$. The study of the above claim is closely related to a the study of properties of the following stochastic process, called the likelihood ration process

$$\Lambda_n(\theta, u) = \log \frac{L_n(\theta + u/\sqrt{n})}{L_n(\theta)}$$

where $u \in \mathcal{R}^d$. The term $\theta + u/\sqrt{n}$ is like a perturbation of magnitude $\frac{1}{\sqrt{n}}$ around θ as for MLE, we have $\hat{\theta}_n - \theta = O_p(\frac{1}{\sqrt{n}})$. Also we have $\hat{u}_n = \arg \max_{\mathcal{R}^d} \Lambda_n(\theta, u) = \sqrt{n}(\hat{\theta}_n - \theta)$. Note that the LLR statistic proposed above could be written using the LLR process as follows

$$\begin{aligned} \Lambda_n &= 2 \left(\log \frac{L_n(\hat{\theta}_n)}{L_n(\theta)} - \log \frac{L_n(\hat{\theta}_n^0)}{L_n(\theta)} \right) \\ &= 2 \left(\log \frac{\sup_{\theta' \in \Theta} L_n(\hat{\theta}'_n)}{L_n(\theta)} - \log \frac{\sup_{\theta' \in L} L_n(\hat{\theta}'_n)}{L_n(\theta)} \right) \\ &= 2 \left(\sup_{u \in \mathcal{R}^d} \Lambda_n(\theta, u) - \sup_{u \in L} \Lambda_n(\theta, u) \right) \end{aligned}$$

Hence, we will examine the first and second term of the above expression separately below. On our way, we also show asymptotic normality of MLE and constrained MLE.

First term First we prove asymptotic normality of MLE. Consider

$$\begin{aligned} \Lambda_n(\theta, u) &= \sum_{i=1}^n (\log P_{\theta+u/\sqrt{n}}(X_i) - \log P_\theta(X_i)) \\ &= \left\langle \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log P_\theta(X_i), u \right\rangle + \frac{1}{2} \left\langle \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log P_\theta(X_i) u, u \right\rangle \end{aligned}$$

Let $Z_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log P_\theta(X_i) \rightarrow_D N(0, I(\theta))$ by central limit theorem and $\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log P_\theta(X_i) = \mathbb{E} \frac{\partial^2}{\partial \theta^2} \log P_\theta(X) = -I(\theta)$ by law of large numbers (asymptotically). Hence

$$\Lambda_n(\theta, u) \approx \langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)u, u \rangle + o_p(1)$$

The above property is called as Local asymptotic normality for a model. The remainder term in the above expression was not defined rigorously. But it will tend to 0 in probability. Much more could be said if we make some more assumptions. We previously saw that $\text{hat}u_n = \sqrt{n}(\hat{\theta}_n - \theta)$. From LAN definition, we have

$$\begin{aligned} \hat{u}_n &= \arg \max_{u \in \mathcal{R}^d} \left(\langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)u, u \rangle \right) \\ \implies \hat{u}_n &= I(\theta)^{-1} Z_n(\theta) \\ &\rightarrow_D I(\theta)^{-1} N(0, I(\theta)) \\ &= I(\theta)^{-1} I(\theta)^{1/2} Z \quad \text{where } Z \text{ is standard normal RV} \\ &= I(\theta)^{-1/2} Z \sim N(0, I(\theta)^{-1}) \end{aligned}$$

Hence we have $\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow_D N(0, I(\theta)^{-1})$. This could be made little more rigorous and made as formal proof. Now consider,

$$\begin{aligned} &\sup_{u \in \mathcal{R}^d} \Lambda_n(\theta, u) \\ &\approx \sup_{u \in \mathcal{R}^d} \left(\langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)u, u \rangle \right) \\ &= \sup_{u \in \mathcal{R}^d} \left(\langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)^{1/2}u, I(\theta)^{1/2}u \rangle \right) \\ &= \sup_{v \in \mathcal{R}^d} \left(\langle Z_n(\theta), I(\theta)^{-1/2}v \rangle - \frac{1}{2} \|v\|^2 \right) \end{aligned}$$

where we define $v = I(\theta)^{1/2}u$. When u runs through \mathcal{R}^d , v also does the same. Hence the problem is reduced to

$$\sup_{v \in \mathcal{R}^d} \left(\langle I(\theta)^{-1/2} Z_n(\theta), v \rangle - \frac{1}{2} \|v\|^2 \right)$$

from which we have $v_{max} = I(\theta)^{-1/2} Z_n(\theta)$ and hence

$$\log \frac{L_n(\hat{\theta}_n)}{L_n(\theta)} \approx \frac{1}{2} \|I(\theta)^{-1/2} Z_n(\theta)\|^2$$

Second term

First we prove asymptotic normality of constrained MLE (Regularized MLE problems may fall under this category). Under H_0 , we have $\hat{\theta}^0 = \arg \max_{\theta \in L} L_n(\theta)$. Similar to the previous case, we have

$$\sqrt{n}(\hat{\theta}_n^0 - \theta) = \hat{u}_n^0 = \arg \max_{u \in L} \Lambda_n(\theta, u)$$

From LAN definition, we have

$$\hat{u}_n^0 = \arg \max_{u \in L} \left(\langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)u, u \rangle \right)$$

If P_L is the orthogonal projection on to L , we have

$$\begin{aligned} \hat{u}_n^0 &= \arg \max_{u \in L} \left(\langle P_L Z_n(\theta), u \rangle - \frac{1}{2} \langle P_L I(\theta) P_L u, u \rangle \right) \\ \implies u_{max} &= (P_L I(\theta) P_L)^{-1} P_L Z_n(\theta) \end{aligned}$$

Hence we have $\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow_D (P_L I(\theta) P_L)^{-1} P_L N(0, I(\theta))$, which could be simplified further. This could be made little more rigorous and made as formal proof. Now consider

$$\begin{aligned} & \sup_{u \in L} \Lambda_n(\theta, u) \\ & \approx \sup_{u \in L} \left(\langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)u, u \rangle \right) \\ & = \sup_{u \in L} \left(\langle Z_n(\theta), u \rangle - \frac{1}{2} \langle I(\theta)^{1/2}u, I(\theta)^{1/2}u \rangle \right) \\ & = \sup_{v \in I(\theta)^{1/2}L} \left(\langle Z_n(\theta), I(\theta)^{-1/2}v \rangle - \frac{1}{2} \|v\|^2 \right) \end{aligned}$$

where we define $v = I(\theta)^{1/2}u$. When u runs through L , v runs through $I(\theta)^{1/2}L = L_1$. Hence the problem is reduced to

$$\sup_{v \in L_1} \left(\langle P_{L_1} I(\theta)^{-1/2} Z_n(\theta), v \rangle - \frac{1}{2} \|v\|^2 \right)$$

from which we have $v_{max} = P_{L_1} I(\theta)^{-1/2} Z_n(\theta)$ and hence

$$\log \frac{L_n(\hat{\theta}_n)}{L_n(\theta)} \approx \frac{1}{2} \|P_{L_1} I(\theta)^{-1/2} Z_n(\theta)\|^2$$

Now that we calculated (approximated) the two terms, we substitute it into LLR we have $\Lambda_n = \|I(\theta)^{-1/2} Z_n(\theta)\|^2 - \|P_{L_1} I(\theta)^{-1/2} Z_n(\theta)\|^2 = \|P_{L_1^\perp} I(\theta)^{-1/2} Z_n(\theta)\|^2$. Hence we could find the distribution of the statistic as follows.

$$\begin{aligned} I(\theta)^{-1/2} Z_n(\theta) & \rightarrow_D I(\theta)^{-1/2} I(\theta)^{1/2} Z = Z \rightarrow_D N(0, I_D) \\ \implies \Lambda_n & \rightarrow_D \|P_{L_1^\perp} Z\|^2 \sim \chi_{d-\dim(L)}^2 \end{aligned}$$

because $L_1 = I(\theta)^{1/2}L$ and $\dim(L_1) = \dim(L)$ which implies $\dim(L_1^\perp) = d - \dim(L)$.

8 Minmax lower bounds for density estimation

First we will look at VC-Theorem.

Theorem 8.1 (VC theorem). *For binary classification and 0/1 loss, we have*

$$\begin{aligned} P \left(\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > \epsilon \right) & \leq 8\mathcal{S}(\mathcal{F}, n) e^{-\frac{n\epsilon^2}{32}} \\ E \left(\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right) & \leq s \sqrt{\frac{\log \mathcal{S}(\mathcal{F}, n) + \log 2}{n}} \end{aligned}$$

and let \hat{f}_n be the classifier chosen by ERM,

$$E(R(\hat{f}_n)) \leq \inf_{f \in \mathcal{F}} R(f) = 4 \sqrt{\frac{VC(\mathcal{F}) \log(n+1) + \log 2}{n}}$$

Similar to the VC theorem, for any estimator \hat{f}_n , we obtain upper bounds of the form $\sup_{f \in \mathcal{F}} E(d(\hat{f}_n, f)) \leq C\eta^{-\gamma}$, $\gamma > 0$. We would like to see if the bounds are tight, i.e., there is no other estimator that is significantly better. That could be done by lower bounds like $\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} E(d(\hat{f}_n, f)) \geq c\eta^{-\gamma}$.

As before, we will be interested in estimator as follows $\sup_{f \in \mathcal{F}} E_f(d(\hat{f}_n, f)) = \sup_{f \in \mathcal{F}} \int d(\hat{f}_n(Z), f) dP_f(Z)$, where P_f is the true distribution of the data (for e.g., in regression $Z = (X, Y)$). In this setting, we would like to obtain bounds like,

$$\mathcal{R}_n^* \stackrel{\text{def}}{=} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} E(d(\hat{f}_n, f)) \geq cs_n.$$

where $c > 0$ and $s_n \rightarrow 0$ as $n \rightarrow \infty$.

Now we will look at some definitions which will be used.

Definition 14. Suppose we show that

$$\liminf_{n \rightarrow \infty} s_n^{-1} \mathcal{R}_n^* \geq c > 0$$

and for a particular estimator \bar{f}_n

$$\limsup_{n \rightarrow \infty} s_n^{-1} \sup_{f \in \mathcal{F}} \mathbf{E}_f (d(\bar{f}_n, f)) \leq C$$

then it would imply, $\limsup_{n \rightarrow \infty} s_n^{-1} R_n^* \leq C$, because, if the risk for best estimator is greater than c and risk for an estimator is lesser than C , then the risk for the best estimator must obviously be lesser than C , too. In this situation, we say that s_n is the optimal rate of convergence for this problem and that \bar{f}_n achieves this rate, in the sense described as follows: Two rates of convergence (Φ_n, Φ'_n) are equivalent, if

$$0 < \liminf_{n \rightarrow \infty} \frac{\Phi_n}{\Phi'_n} \leq \limsup_{n \rightarrow \infty} \frac{\Phi_n}{\Phi'_n} < \infty$$

Instead of directly bounding the expected performance, we will prove stronger bounds of the form,

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P(d(\hat{f}_n, f) \geq s_n) \geq c > 0.$$

Then by Markov's inequality,

$$\begin{aligned} P_f(d(\hat{f}_n, f) \geq s_n) &\leq \frac{\mathbf{E}_f(d(\hat{f}_n, f))}{s_n} \\ \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbf{E}_f(d(\hat{f}_n, f)) &\geq \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f(d(\hat{f}_n, f) \geq s_n) \geq cs_n \end{aligned}$$

- Reduce the infinite class \mathcal{F} with finite $\{f_0, \dots, f_M\} \subset \mathcal{F}$. Then we will have $\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f(d(\hat{f}_n, f) \geq s_n) \geq \inf_{\hat{f}_n} \sup_{f \in \{f_0, \dots, f_M\}} P_f(d(\hat{f}_n, f) \geq s_n)$ (The idea is to choose a finite collection of models such that the resulting problem is as hard as original, i.e., need to make the lower bound tight).
- Reduce the problem to hypothesis testing, i.e.,

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f(d(\hat{f}_n, f) \geq s_n) \geq \inf_{\hat{f}_n} \sup_{j \in \{0, \dots, M\}} P_{f_j}(\hat{h}_n(Z) \neq j \geq s_n)$$

where $\hat{h}_n : Z \rightarrow \{0, \dots, M\}$ is the set of all measurable test functions and $P_{f_j}(\hat{h}_n \neq j)$ denotes the probability that after observing the data, the test infers the wrong hypothesis. Below, we give a Lemma which gives insights about how to construct the test.

Definition 15. Suppose $d(\cdot, \cdot)$ is a semi metric. And we have constructed f_0, \dots, f_M , such that $d(f_j, f_k) \geq 2s_n, \forall j \neq k$. Take any estimator \hat{f}_n and define the test $\Phi^* \circ \hat{f}_n : Z \rightarrow \{0, \dots, M\}$ as $\Phi^*(\hat{f}_n) = \arg \min_j d(\hat{f}_n, f_j)$, then $\Phi^*(\hat{f}_n) \neq j \implies d(\hat{f}_n, f_j) \geq s_n$.

From previous lemma, note that

$$\begin{aligned} P_{f_j}(d(\hat{f}_n, f_j) \geq s_n) &\geq P_{f_j}(\Phi^*(\hat{f}_n) \neq j) \\ \inf_{\hat{f}_n} \max_{f \in \{f_0, \dots, f_M\}} P_f(d(\hat{f}_n, f) \geq s_n) &\geq \inf_{\hat{f}_n} \max_{j \in \{0, \dots, M\}} P_{f_j}(\Phi^*(\hat{f}_n) \neq j) \\ &\geq \inf_{\hat{h}_n} \max_{j \in \{0, \dots, M\}} P_j(\hat{h}_n \neq j) \stackrel{\text{def}}{=} P_{e,M} \end{aligned}$$

where the last step follows because we replace the class of test defined by $\Phi^*(\hat{f}_n)$ by a large class of ALL possible test and inf is taken over the larger class is smaller. Now, we need to find lower bounds on $P_{e,M}$ by constructing $\{f_0, \dots, f_M\}$ such that distance between each is greater than $2s_n$ and $P_{e,M} \geq c > 0$. These are contradicting- condition 1 requires f_j, f_k be far apart, but 2 requires them to be closer so that they are indistinguishable and hence the prob of error is bounded away from 0.

Now for $M = 1$, and $\hat{h}_n(Z) = I_A(Z)$, for any subset A , we have

$$\begin{aligned}
P_{e,1} &= \inf_{\hat{h}_n} \max_{j \in \{0, \dots, M\}} P_j(\hat{h}_n \neq j) \\
&\geq \inf_{\hat{h}_n} \left(\frac{1}{2} P_0(\hat{h}_n \neq 0) + P_1(\hat{h}_n \neq 1) \right) \\
&= \frac{1}{2} \inf_A P_0(I_A(Z) \neq 0) + P_1(I_A(Z) \neq 1) \\
&= \frac{1}{2} \inf_A P_0(A) + P_1(A^c) \\
&= \frac{1}{2} \inf_A (1 - (P_0(A) + P_1(A))) \\
&= \frac{1}{2} (1 - d_{TV}(P_0, P_1))
\end{aligned}$$

We see that if P_0 is close to P_1 , then $d_{TV}(P_0, P_1)$ is small and the probability of error $P_{e,1}$ is large. But this is difficult to work with. Hence, we use the fact that $1 - d_{TV}(P_0, P_1) \geq \frac{1}{2} \exp -K(P_1, P_0)$.

Hence, we choose f_0, f_1 such that $K(P_1, P_0) \leq \alpha$, the $P_{e,1}$ is bounded away from 0 and we get the required lower bounds.

Often it turns out that reducing the initial problem to binary testing does not always work. sometime we need $M \rightarrow \infty$ as $n \rightarrow \infty$.

Theorem 8.2. *let $M \geq 2$ and $\{f_0, \dots, f_M\} \in \mathcal{F}$ be such that,*

- $d(f_j, f_k) \geq 2s_n$ d is semi-distance.
- $\frac{1}{M} \sum_{i=1}^M K(P_j, P_0) \leq \alpha \log M$, with $0 < \alpha < 1/8$. Then,

$$\begin{aligned}
\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f(d(\hat{f}_n, f) \geq s_n) &\geq \inf_{\hat{f}_n} \max_{j \in \{0, \dots, M\}} P_j(d(\hat{f}_n, f_j) \geq s_n) \\
&\geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log M}} \right)
\end{aligned}$$

9 Structural assumptions made for high dimensional estimation

In high dimensional setting, if we do not make an assumption, the variance of our estimate is very high. Hence, we kind of make some assumptions about the models and allow for lower variance in estimation, such that the overall MSE (bias-squared + variance) is still much smaller than if we did not make any such assumptions. In that settings comes L1 (sparsity), L2 regularization or some assumptions on the dependency structure (as in graphical models) so that high dimensional estimation problem is broken down into many simpler low dimensional problems. These could be mixed up, i.e., we can make dependency and L1 etc (graphical lasso).

10 Conditional expectations and martingales

Definition 16 (Conditional Expectation). Let (Ω, \mathcal{F}, P) be a probability space, and let $\mathcal{C} \subset \mathcal{F}$ be a sub σ -field. Let X be a r.v. whose mean is defined. Denote by $E(X|\mathcal{C})$ for any function $h : \Omega \rightarrow \mathbb{R}$ that is \mathcal{C} measurable and that satisfies

$$\int_C h dP = \int_C X dP, \quad \forall C \in \mathcal{C}.$$

Such a function h is called as conditional expectation of X given \mathcal{C} . If Y is some random variable and if $\mathcal{C} = \sigma(Y)$, we use the notation $E(X|Y)$ for conditional expectation of X given Y .

Definition 17 (Filtration). Given a measurable space (Ω, \mathcal{F}) , a filtration is a sequence of sigma algebras $\{\mathcal{F}_t\}_{t \geq 0}$ with $\mathcal{F}_t \subset \mathcal{F}$ for each t and such that $t_1 \leq t_2 \implies \mathcal{F}_{t_1} \subset \mathcal{F}_{t_2}$.

A σ -algebra defines the set of events that can be measured, which in a probability context is equivalent to events that can be discriminated, or questions that can be answered at time t . Therefore a filtration is often used to represent

the change in the set of events that can be measured, through gain or loss of information. A typical example is in mathematical finance, where a filtration represents the information available at each time t , and is more and more precise (the set of measurable events is staying the same or increasing) as information from the present becomes available.

Definition 18 (Martingale). Let $\{\mathcal{F}_n\}_{n=1}^\infty$ be a filtration on the probability space (Ω, \mathcal{F}, P) . If $X_n : \Omega \rightarrow \mathbb{R}$ is \mathcal{F}_n -measurable for every n , then we say that $\{X_n\}_{n=1}^\infty$ is *adapted* to the filtration. If $\{X_n\}_{n=1}^\infty$ is adapted to filtration $\{\mathcal{F}_n\}_{n=1}^\infty$ and if $\mathbb{E}|X_n| < \infty$ for all n and $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all n , then we say that $\{X_n\}_{n=1}^\infty$ is a martingale relative to the filtration.

If $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$, then $\{X_n\}_{n=1}^\infty$ is called sub martingale. If $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$, then $\{X_n\}_{n=1}^\infty$ is called super martingale.

- Let $\{X_n\}_{n=1}^\infty$ be a martingale relative to the filtration $\{\mathcal{F}_n\}_{n=1}^\infty$. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a convex function such that $\mathbb{E}(\phi(X))$ is finite for all n . Define $Y_n = \phi(X_n)$. Then

$$\begin{aligned} \mathbb{E}(Y_{n+1}|\mathcal{F}_n) &= \mathbb{E}(\phi(X_{n+1})|\mathcal{F}_n) \\ &\geq \phi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \\ &= \phi(X_n) = Y_n \end{aligned}$$

Hence $\{Y_n\}_{n=1}^\infty$ is a sub martingale relative to the filtration $\{\mathcal{F}_n\}_{n=1}^\infty$.

11 Entropy of Metric Spaces

Theorem 11.1 (GC theorem). Let $Z^{(i)}, i = 1, \dots, n$ be iid with CDF $F(t) = P(Z \leq t)$. Define Empirical CDF $\hat{F}_n(t) = \frac{1}{n} \mathcal{I}(Z^{(i)} \leq t)$. Then

$$\lim_{n \rightarrow \infty} P\left(\sup_{t \in \mathcal{R}} |\hat{F}_n(t) - F(t)| > \epsilon\right) \rightarrow 0$$

Let P be a measure on $(\mathcal{X}, \mathcal{A})$ and $L_p(Q) = \{g : \mathcal{X} \rightarrow \mathbb{R} : \|g\|_{p,Q}^p = \int |g|^p dQ < \infty\}, 1 \leq p \leq \infty$. Distance between two function in this space is denoted by $\|g_1 - g_2\|_{p,Q}^p$

Definition 19 (Entropy for $L_p(Q)$). Consider for each $\delta > 0$, a collection of functions $g_1 \cdots g_N$, such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \dots, N\}$, such that

$$\|g - g_j\|_{p,Q} \leq \delta$$

Let $N_p(\delta, \mathcal{G}, Q)$ be the smallest value of N for which such a covering by balls with radius δ and centers g_1, \dots, g_N exists. Its called the δ -covering number and $H_p(\delta, \mathcal{G}, Q) = \log N_p(\delta, \mathcal{G}, Q)$ is called the δ -entropy number

Definition 20 (Entropy with bracketing for $L_p(Q)$). Let $N_{p,B}(\delta, \mathcal{G}, Q)$ be the smallest value of N for which there exists pairs of functions $\{[g_j^L, g_j^U]\}$ such that $\|g_j^L - g_j^U\|_{p,Q} \leq \delta \quad \forall j$ and such that for each $g \in \mathcal{G}$, there exists a $j = j(g)$ such that

$$g_j^L \leq g \leq g_j^U$$

Then $H_{p,B}(\delta, \mathcal{G}, Q) = \log N_{p,B}(\delta, \mathcal{G}, Q)$ is called the δ -entropy with bracketing for \mathcal{G} .

Definition 21 (Entropy with supremum norm). Let $N_\infty(\delta, \mathcal{G})$ be the smallest value of N for which there exists pairs of functions $\{g_j\}_{j=1}^N$ with

$$\sup_{g \in \mathcal{G}} \min_{j=1, \dots, N} |g - g_j|_\infty \leq \delta$$

Then $H_\infty(\delta, \mathcal{G}) = \log N_\infty(\delta, \mathcal{G})$ is called the δ -entropy of \mathcal{G} with supremum norm.

Definition 22 (A condition for ULLN). Suppose that $H_{1,B}(\delta, \mathcal{G}, P) < \infty$ for all $\delta > 0$, then \mathcal{G} satisfies the ULLN. This is a much stronger which could be relaxed.

Definition 23 (ULLN based on chaining). The function $G = \sup_{g \in \mathcal{G}} |g|$ is called the envelope of \mathcal{G} . If $G \in L_1(P)$ (infact prev defnion also implies this) and $\frac{1}{n} H_1(\delta, \mathcal{G}, P) \rightarrow_p 0$ for all $\delta > 0$, then ULLN holds good.

12 Rate of convergence of MLE via Empirical process

Let X_1, \dots, X_n be i.i.d. samples from unknown density function p_0 and let \mathcal{P} be the collection of densities (which also includes p_0). Then, $\hat{p}_n(X_1, \dots, X_n)$ is a MLE for p_0 if for any $p \in \mathcal{P}$, we have $\prod_{i=1}^n \hat{p}_n(X_i) \geq \prod_{i=1}^n p(X_i)$. In order to present an unified view of finite and infinite dimensional parameters, we work with densities rather than the parameters. We consider the space \mathcal{P} endowed with the hellinger metric, $h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu$ as (pseudo)metric space (\mathcal{P}, h) . The hellinger distance(h) is bounded above by $\sqrt{2}$. This is one of the nicest properties of this distance over standard distance like KL distance which is potentially unbounded. We say convergence rate of $p_n \rightarrow p$ is $O(\epsilon_n)$ if $h(p_n, p) \rightarrow 0$ and $h(p_n, p) = O_p(\epsilon_n)$.

Non-parametric MLEs: We consider non-parametric class of functions \mathcal{P} as it its known to increase robustness. But the estimator in such a case may not be as powerful as that in the parametric case. Its known that empirical measure is the MLE if we take \mathcal{P} to be the set of all possible densities. But, for the case of continuous densities, this MLE is zero almost everywhere. Hence, we see that it does not converge to the true density in the way we want it to and hence not consistent. So, we do some relaxations on (i) definition of MLE and (ii) the size of class of densities considered.

Relaxation 1: Let X_1, \dots, X_n be iid samples from p_0 and let $\eta_n \rightarrow 0$ be a positive real sequence. Then, $\hat{p}_n(X_1, \dots, X_n)$ is an η_n -MLE of p_0 if for any $p \in \mathcal{P}$ if $\prod_{i=1}^n \hat{p}_n(X_i) \geq \prod_{i=1}^n p(X_i) - \eta_n$. The motivation is if we work with set of all densities over infinite dimensional spaces, its possible that there is no exact MLE.

Relaxation 2: We define a quantity called as ϵ -entropy below.

Definition 24 (Entropy for $L_P(\mu)$ as Hellinger distance). Consider for each $\epsilon > 0$, a collection of functions $p_1 \cdots p_N$, such that for each $p \in \mathcal{P}$, there is a $j = j(p) \in \{1, \dots, N\}$, such that

$$h(p, p_j)_{p, \mu} \leq \epsilon$$

Let $N_p(\epsilon, \mathcal{P}, h)$ be the smallest value of N for which such a covering by balls with radius ϵ and centers p_1, \dots, p_N exists. Its called the ϵ -covering number and $H_p(\epsilon, \mathcal{P}, h) = \log N_p(\epsilon, \mathcal{P}, h)$ is called the ϵ -entropy number

We are interested in $H_p(\epsilon, \mathcal{P}, h)$ for the following reason. Consider random variables $\{Z_i\}$ which are exponentially bounded, i.e., $P(Z_i \geq a) \leq \exp(-f(a))$, for some increasing function f . Then $P(\max_{i=1, \dots, N} Z_i \geq a) \leq N \exp(-f(a)) = \exp(\log N - f(a))$. We see that the interesting case is when $\log N < f(a)$ and hence we are interested in the $\log N$.

Now define $Z_i = Z(p_i) = \prod_{j=1}^n \frac{p_i(X_j)}{p_0(X_j)}$. In that case, $P(Z_i \geq a)$ means p_i is in some sense not very close to p_0 . Then we can also ask what is the probability of the event that the maximum of all p_i is also not very close to p_0 ? From the fact that $P(\max_{i=1, \dots, N} Z_i \geq a) \leq N \exp(-f(a)) = \exp(\log N - f(a))$ and since we know that $Z(\hat{p}_n) \geq 1$, we can get probability that \hat{P}_n is among the densities which are not very close to p_0 . By selecting a class of densities which is not too large, we can use the exponential inequality. This leaves us with the task of formulating suitable class of functions and calculating their entropies which could be useful. Another suitable variation of the ϵ -entropy defined above is the ϵ -entropy with bracketing, which is easier to calculate and also use.

Definition 25 (Entropy with bracketing for $L_P(Q)$ as hellinger metric). Let $N_{p,B}(\delta, \mathcal{P}, h)$ be the smallest value of N for which there exists pairs of functions $\{[p_j^L, p_j^U]\}$ such that $h(p_j^L, p_j^U)_{p, \mu} \leq \delta \quad \forall j$ and such that for each $p \in \mathcal{P}$, there exists a $j = j(p)$ such that

$$p_j^L \leq p \leq p_j^U$$

Then $H_{p,B}(\delta, \mathcal{P}, h) = \log N_{p,B}(\delta, \mathcal{P}, h)$ is called the δ -entropy with bracketing for \mathcal{P} .

Now we are ready to see the relation between the rate of convergence of MLE using these entropies.

Theorem 12.1. *Let \mathcal{P} be a class of densities containing p_0 and let $H_{p,B}(\delta, \mathcal{P}, h)$ be the Hellinger δ -entropy with bracketing for \mathcal{P} . Then there exists constants c_1, c_2, c_3, c_4 such that if*

$$\int_{\delta^2/2^8}^{\sqrt{2}\delta} H_{p,B}^{1/2}(u/c_3, \mathcal{P}, h) du \leq c_4 n^{1/2} \delta^2$$

then for sufficiently large n , we have

$$P \left(\sup_{\{p \in \mathcal{P}: \|p^{1/2} - p_0^{1/2}\|_2 \geq \delta\}} \prod_{i=1}^n \frac{p(X_i)}{p_0(Y_i)} \geq \exp(c_2 n \delta^2) \right) \leq 4 \exp(-c_2 n \delta^2)$$

The essence of this theorem is that the likelihood ratio is uniformly exponentially small outside a hellinger ball of radius δ around the true density. Thus, since we know that $\prod_{i=1}^n \frac{\hat{p}(X_i)}{p_0(Y_i)} \geq 1 > \exp(-c_1 n \delta^2)$, we can conclude that for large n , \hat{p}_n is likely to lie inside a hellinger ball of radius δ around p_0 .

Theorem 12.2. *Let $\delta_n \rightarrow 0$ be a sequence such that the integral condition in the above theorem is satisfied with $\delta = \delta_n$ for each n . Let \hat{p}_n be an η_n -MLE for p_0 , with $\eta_n \leq c_1 \delta_n^2$. Then for large n , we have*

$$P(h(\hat{p}_n, p_0) \geq \delta_n) \leq 4 \exp(-c_2 n \delta_n^2)$$

Thus the rates of convergence is $O(\delta_n)$

Proof. By the definition of η_n -MLE, we have

$$\{\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \delta_n\} \subset \left\{ \sup_{\{p \in \mathcal{P}: \|p^{1/2} - p_0^{1/2}\|_2 \geq \delta_n\}} \prod_{i=1}^n \frac{p(X_i)}{p_0(Y_i)} \geq \exp(-n\eta_n) \right\}$$

Since $\exp(-n\eta_n) \geq \exp(-nc_2 \delta_n)$, by applying the result of previous theorem, we have

$$P(\|\hat{p}_n^{1/2}, -p_0^{1/2}\|_2 \geq \delta_n) \leq 4 \exp(-c_2 n \delta_n^2)$$

□

The combination of the above two theorems says that the convergence of MLE is basically determined by

$$\int_{\delta^2}^{\delta} H_{p,B}^{1/2}(\delta, \mathcal{P}, h) = n^{1/2} \delta^2.$$

This theorem is the most basic theorem in this line of research. Though this result is nice, it requires an envelope condition on the densities. i.e., $H_{p,B}(u, \mathcal{P}, h) \leq \infty$ for all u which means that we need, $\int G(x) dx < \infty$, where $G = \sup_G |g|$. Hence, we try to relax this condition next.

Define $\bar{\mathcal{P}}^{1/2} = \left\{ \left(\frac{p+p_0}{2} \right)^{1/2} : p \in \mathcal{P} \right\}$ and let \hat{p}_n be a maximum likelihood estimator. Let $H_{p,B}(u, \mathcal{P}, \|\cdot\|_2)$ be the δ -entropy with bracketing using the L_2 norm.

Theorem 12.3. *Suppose that δ_n is a sequence satisfying*

$$\Phi(\delta_n) \stackrel{\text{def}}{=} \delta_n \vee \int_{\delta_n^2/2^{13}}^{\delta_n} H_{p,B}^{1/2}(u, \bar{\mathcal{P}}^{1/2}, \|\cdot\|_2) du \leq c_1^{-1} n^{1/2} \delta_n^2$$

for some constant c_1 . Then provided that $\Phi(\delta)/\delta^2$ is non-increasing,

$$P(h(\hat{p}_n, p_0) \geq \delta_n) \leq c_2 \exp\left(-\frac{n\delta_n^2}{c_2^2}\right)$$

Since we have $H_{p,B}^{1/2}(u, \bar{\mathcal{P}}^{1/2}, \|\cdot\|_2) = H_{p,B}^{1/2}(u, \bar{\mathcal{P}}, h)$, and $h\left(\frac{p+p_0}{2}, p_0\right) \leq \frac{1}{\sqrt{2}} h(p, p_0) \leq 2\sqrt{2} h\left(\frac{p+p_0}{2}, p_0\right)$, the above theorem is a restatement of the previous theorem. But the advantage is we use L_2 as distance metric.

Now, suppose that \mathcal{P} is a convex class of densities and define $\mathcal{P}^* = \left\{ \frac{2pp_0}{p+p_0} : p \in \mathcal{P} \right\}$.

Theorem 12.4. *Suppose that δ_n is a sequence satisfying*

$$\Phi(\delta_n) \stackrel{\text{def}}{=} \delta_n \vee \int_{\delta_n^2/c_1}^{\delta_n} H_{p,B}^{1/2}(u, \mathcal{P}^*, \|\cdot\|_2) du \leq c_1^{-1} n^{1/2} \delta_n^2$$

for some constant c_1 . Then provided that $\Phi(\delta)/\delta^2$ is non-increasing,

$$P(h(\hat{p}_n, p_0) \geq \delta_n) \leq c_2 \exp\left(-\frac{n\delta_n^2}{c_2^2}\right)$$

for some constant c_2 .

The advantage of this theorem is this does not entail the envelope condition. We look at one illustrative example, which is slightly more general.

Let $\mathcal{P} = \{p : [0, 1] \rightarrow \mathcal{R}_+, \int p d\mu = 1, |p'| \leq M < \infty\}$. This makes no assumption on the square root densities. Hence there is no way to bound the bracketing entropy of \mathcal{P} under hellinger distance. However, subjected some conditions on p_0 , the class $\mathcal{P}^* = \{\frac{2pp_0}{p+p_0} : p \in \mathcal{P}\}$ is such that $H_{p,B}(\delta, \mathcal{P}^*, \|\cdot\|_2) < A\delta^{-1}$ for some constant A . Then by applying the previous theorem, we can show that $h(\hat{p}_n, p_0) = O_p(n^{-1/3})$.

Using this definition of δ -entropy, one can extract the parametric rates as $O(n^{-1/2} \log n)$. One can also extract the exact parametric rates by a something called as local entropy with bracketing, i.e.,

$$\int_{\frac{\delta^2}{2^8}}^{\sqrt{2}\delta} H_{p,B}^{1/2}(u/c_3, \mathcal{P}(\delta), h) du \leq c_4 n^{1/2} \delta^2$$

where $\mathcal{P}(\delta) = \{p \in \mathcal{P} : h(p, p_0) \leq \delta\}$.

Sieves:

Another way to deal with large parameters is through the concept of sieves.

Definition 26. Let $\epsilon_n \rightarrow 0$ be a real sequence and let (U, d) be a metric space, with sequence of subsets $U_n \subset U$. Suppose for each n , U_n is an ϵ_n -net for U ; then we call U_n and ϵ_n sieve.

Definition 27. Given the true density p_0 and another density p , we set

$$\rho_\alpha(p_0, p) = \begin{cases} \frac{1}{\alpha} \left(\mathbf{E} \left(\frac{p_0}{p} \right)^\alpha - 1 \right) & \text{for } \alpha \in [-1, 1] - 0 \\ \mathbf{E}_{p_0} \log \left(\frac{p_0}{p} \right) & \alpha = 0 \end{cases}$$

Note that $\alpha = -1/2$ gives squared hellinger distance, and $\alpha = 0$ gives the KL divergence.

Theorem 12.5. Let \mathcal{P} be a class of densities and let \mathcal{P}_n be a sequence of subsets of \mathcal{P} such that

$$\inf_{p \in \mathcal{P}_n} \rho_\alpha(p_0, p) \leq \epsilon_n < \frac{1}{\alpha}$$

for some $\alpha \in (0, 1]$. With δ_n as the smallest value of δ , define

$$\delta_n^* = \begin{cases} \delta_n & \text{if } \epsilon_n < \frac{c_1 \delta_n^2}{4} \\ (4\epsilon_n/c_1)^{1/2} & \text{otherwise} \end{cases}$$

Then if \hat{p}_n is an η_n -MLE, where $\eta_n < \frac{c_1(\delta_n^*)^2}{2}$,

$$P(h(\hat{p}_n, p_0) \geq \delta_n) \leq C \exp(Dn(\delta_n^*)^2)$$

for some constant C,D.

13 Influence functions and efficient estimation

In this section we consider estimation in semiparametric models of the form $\{p_\theta(z), \theta \in \Theta \subset \mathbb{R}^p\}$ with $\theta = (\beta^\top, \eta^\top)$, with $\beta \in \mathbb{R}^q$ and $\eta \in \mathbb{R}^r$ and $p = q + r$. β is parameter of interest and η is nuisance parameter. The parametric component is also defined as $\beta(\theta)$. From a previous section, we saw that a reasonable estimator of the parameter β must have the form $n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^n \psi(x_i) + o_p(1)$ with $\mathbf{E}(\psi(x)) = 0^{q \times 1}$ and $\mathbf{E}(\psi\psi^\top) < \infty$ and non-singular. Let $S_{\theta_0}(x) = \frac{\partial \log p_\theta(x)}{\partial \theta}$ at $\theta = \theta_0$ be the score function for a single observation. This could be partitioned into $\mathbb{R}^{q+r} \ni S_{\theta_0}(x) = (S_{\beta_0}^\top(x), S_{\eta_0}^\top(x))^\top$.

Theorem 13.1. Let the parameter of interest $\beta(\theta)$ be a q dimensional function of the p dimensional parameter θ such that $\Gamma^{q \times p}(\theta) = \frac{\partial \beta(\theta)}{\partial \theta^\top}$, a $q \times p$ dimensional matrix of partial derivatives with rank q . Let $\hat{\beta}_n$ be an asymptotically linear estimator with influence function $\psi(x)$ such that covariance matrix exists. Then, under proper regularity conditions,

$$\mathbf{E}\{\psi(x)S_{\theta_0}^\top(x)\} = \Gamma(\theta_0)$$

In the case when $\theta = (\beta, \eta)$, we also have

$$\begin{aligned} E\{\psi(x)S_{\beta_0}^\top(x)\} &= I^{q \times q} \\ E\{\psi(x)S_{\eta_0}^\top(x)\} &= 0^{q \times r} \end{aligned}$$

m -estimators:

Consider a $p \times 1$ dimensional function of x and θ , $m(z, \theta)$ such that

$$E_\theta\{m(x, \theta)\} = 0^{p \times 1}$$

with $E_\theta\{m^\top(x, \theta)m(x, \theta)\} < \infty$ and $E_\theta\{m(x, \theta)m^\top(x, \theta)\}$ positive definite. The m -estimator is defined as the solution the equation

$$\sum_{i=1}^n \{m(x_i, \theta)\} = 0^{p \times 1}$$

for $x_1, \dots, x_n \stackrel{\text{iid}}{\sim} p_\theta$. The MLE is a m -estimator with $m(x, \theta) = S_\theta(x)$. One of the condition needed for proving consistency and asymptotic normality is $n^{-1} \sum_{i=1}^n \frac{\partial m(x_i, \theta)}{\partial \theta^\top} \rightarrow_p E_{\theta_0} \left\{ \frac{\partial m(x, \theta)}{\partial \theta^\top} \right\}$ uniformly in θ , in a small neighbourhood of θ_0 , with $E_{\theta_0} \left\{ \frac{\partial m(x, \theta)}{\partial \theta^\top} \right\} \in \mathbb{R}^{p \times p}$ and non-singular, This would be satisfied if the sample paths of $\frac{\partial m(x, \theta)}{\partial \theta^\top}$ are continuous in θ about θ_0 almost surely and

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left| \frac{\partial m(x, \theta)}{\partial \theta^\top} \right| \leq g(x), \quad E\{g(x)\} < \infty$$

where $\mathcal{N}(\theta_0)$ is a small neighbourhood of θ_0 . Under these conditions, we would ve a regular asymptptically linear estimator $\hat{\theta}_n$, given by

$$n^{1/2}(\hat{\theta}_n - \theta_0) = - \left(E_{\theta_0} \left(\frac{\partial m(x, \theta)}{\partial \theta^\top} \right) \right)^{-1} \left(n^{-1/2} \sum_{i=1}^n m(x_i, \theta_0) \right) + o_p(1).$$

Hence, we can see the influence function is given by $\psi(x_i) = - \left(E_{\theta_0} \left(\frac{\partial m(x, \theta)}{\partial \theta^\top} \right) \right)^{-1} m(x_i, \theta_0)$ and

$$n^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D N \left(0, \left(E_{\theta_0} \left(\frac{\partial m(x, \theta)}{\partial \theta^\top} \right) \right)^{-1} E\{m(x, \theta_0)m^\top(x, \theta_0)\} \left(E_{\theta_0} \left(\frac{\partial m(x, \theta)}{\partial \theta^\top} \right) \right)^{-\top} \right).$$

This influence function could be partitioned as $\psi_{\hat{\theta}_n}(x_i) = (\psi_{\hat{\beta}_n}(x_i), \psi_{\hat{\eta}_n}(x_i))$. Taking the covariance of influence and score functions, we see that

$$E\{\psi_{\hat{\theta}_n}(x_i)S_{\theta_0}(x_i)\} = - \left(E_{\theta_0} \left(\frac{\partial m(x, \theta)}{\partial \theta^\top} \right) \right)^{-1} E\{m(x_i, \theta_0)S_{\theta_0}^\top(x)\}$$

which will be a $I^{(q+r) \times (q+r)}$ matrix. with $E\{\psi_{\hat{\beta}_n}(x_i)S_{\beta_0}(x_i)\} = I^{q \times q}$ and $E\{\psi_{\hat{\eta}_n}(x_i)S_{\eta_0}(x_i)\} = 0^{q \times r}$. Now, we proceed to give a geometric interpretation of the influence function.

Let \mathcal{H} be a hilbert space with q dimensional measurable functions with mean zero and finite variace. Let $\mathcal{L} \subset \mathcal{H}$ be a finite dimensional linear subspace spanned by the p -dimensional score vector $S_{\theta_0}(x)$, i.e., it will be the set of all q -dimensional mean zero random vectors consisting of $B^{q \times p}S_{\theta_0}(x)$, for all $q \times p$ matrices B .

When $\theta = (\beta^\top, \eta^\top)^\top$, the space $B^{q \times r}S_{\eta_0}(x)$ spanned by the nuisance tangent space Λ , from which we note that the q -dimensional influence function $\psi_{\hat{\beta}_n}(x)$ is orthogonal to the nuisance tangent space. It also should satisfy $E\{\psi_{\hat{\beta}_n}(x_i)S_{\beta_0}(x_i)\} = I^{q \times q}$. A question to ask is if the converse is true, i.e., for any element of the Hilbert space satisfying the above conditions, does there exist an RAL estimator with that corresponding influence function for the parameter β . The answer turns out to be yes (under some technical conditions). It also gives us a way for constructing estimators.

Given $\psi(x)$, a q -dimensional measurable function with zero mean and finite variance, define $m(x, \beta, \eta) = \psi(x) - E_{\beta, \eta}\{\psi(x)\}$. It could be shown that the solution to the equation $\sum_{i=1}^n m(x_i, \beta, \hat{\eta}_n(\beta)) = 0$ will be an asymptically

linear estimator for β with influence function $\psi(x)$ (where $\hat{\eta}_n$ is a \sqrt{n} consistent estimator for η). This is mostly a theoretical study for proving converse. It may or may not be of use in practice.

The usefulness of the above fact is that it could be used for comparing RAL estimators for β by looking at the asymptotic variance as we have $n^{-1/2}(\hat{\beta}_n - \beta_0) \rightarrow_D N(0, E(\psi\psi^\top))$, i.e., asymptotic variance of an RAL estimator is the variance of the influence function. But the influence functions could be viewed as members of a subspace of the Hilbert space. In a Hilbert space distance from origin is equal to the variance of the element. Hence the design of best estimators is equivalent to search for the element in the subspace of the influence functions that has the shortest distance to the origin.

Constructing efficient influence functions

In order to discuss this, we first need some definitions.

Definition 28. We say that $M \oplus N$ is a direct sum of two linear subspaces $M, N \subset \mathcal{H}$ if $M \oplus N$ is a linear subspace of \mathcal{H} and if every element $x \in M \oplus N$ has a unique representation of the form $x = m + n$, $m \in M$ and $n \in N$. Speaking non-rigorously, one could always represent a Hilbert space $\mathcal{H} = M \oplus M^\perp$, where M^\perp is an orthogonal complement of a linear subspace M .

According to the above definition, note that the influence function belongs to Λ^\perp , where Λ is nuisance tangent space.

Definition 29. Consider an element $h \in \mathcal{H}$. By projection theorem, there exists an element $a \in \Lambda$, such that $\|h - a_0\|$ has minimum norm and a_0 must uniquely satisfy the relationship $\langle h - a_0, a \rangle = 0$ for all $a \in \Lambda$. The element a_0 is referred as projection of h on to the space Λ and is denoted by $\Pi(h|\Lambda)$. The quantity $h - a_0$ is called the residual and $h - a_0 = \Pi(h|\Lambda^\perp)$.

As we have the space $\Lambda = \{B^{q \times r} S_{\eta_0}(x)\}$ for all $B^{q \times r}$ matrix, the elements orthogonal to this set of nuisance tangent space will be of the form $h - \Pi(h|\Lambda)$ for all $h \in \mathcal{H}$, where by simple calculation, we have $\Pi(h|\Lambda) = E(h S_\eta^\top) \{E(S_\eta S_\eta^\top)\}^{-1} S_{\eta_0}(x)$. It is also easy to show that the tangent space (\mathcal{F}) will be a direct sum of nuisance tangent space and parameter-of-interest tangent space.

Definition 30 (Linear variety). A linear variety is a translation of a linear subspace away from the origin, i.e., a linear variety V can be written as $V = x_0 + M$, where $x_0 \in \mathcal{H}$ and $x_0 \notin M$, $\|x_0\| \neq 0$ and M is a linear subspace.

The set of all influence functions, namely the elements of \mathcal{H} (and which satisfies the two conditions) will form a variety $\psi^*(x) + \mathcal{F}^\perp$.

Theorem 13.2. *The efficient influence function is given by*

$$\psi_{eff}(x) = \psi^*(x) - \Pi(\psi^*(x)|\mathcal{F}^\perp) = \Pi(\psi^*(x)|\mathcal{F})$$

where $\psi^*(x)$ is an arbitrary influence function and \mathcal{F} is the tangent space. One could also write explicitly the efficient influence function as follows

$$\psi_{eff}(x) = \Gamma(\theta_0) I^{-1}(\theta_0) S_{\theta_0}(x)$$

Similarly we have $S_{eff, \theta_0}(x) = S_{\beta_0}(x) - \Pi(S_{\beta_0}(x)|\Lambda)$, where $\Pi(S_{\beta_0}(x)|\Lambda) = E(S_\beta S_\eta^\top) \{E(S_\eta S_\eta^\top)\}^{-1} S_{\eta_0}(x)$. When the parameter space could be partitioned into β, η , we have $\psi_{eff}(x) = \{E(S_{eff} S_{eff}^\top)\}^{-1} \{S_{eff, \theta_0}(x)\}$. From this we can see the well known result for minimum variance of RAL estimator with nuisance parameter as $\{I_{\beta\beta} - I_{\beta\eta} I_{\eta\eta}^{-1} I_{\eta\beta}^\top\}^{-1}$, where the I s are the corresponding information sub-matrices. This view is particularly helpful when we are dealing with infinite-dimensional nuisance parameters.

References

Most of the material in this note is taken from the following books

1. A. W. van der Vaart, *Asymptotic Statistics*, 1998.
2. A. B. Tsybakov, *Introduction to Non-parametric Estimation*, 2009.
3. S. van der Geer, *Empirical process in M-Estimation*, 2000.
4. A. DasGupta, *Asymptotic Theory of Statistics and Probability*, 2008.