Theory of Statistics II (Fall 2016)

J.P.Kim

Dept. of Statistics

Finally modified at October 6, 2016

Preface & Disclaimer

This note is a summary of the lecture Theory of Statistics II (326.522) held at Seoul National University, Fall 2016. Lecturer was B.U.Park, and the note was summarized by J.P.Kim, who is a Ph.D student. There are few textbooks and references in this course. Contents and corresponding references are following.

- Asymptotic Approximations. Reference: *Mathematical Statistics: Basic ideas and selected topics, Vol. I., 2nd edition, P.Bickel & K.Doksum, 2007.*
- Weak Convergence. Reference: *Convergence of Probability Measures, P.Billingsley, 1999.*
- Empirical Processes. Reference: *Empirical Processes in M-estimation, S.A. van de Geer, 2000.*

Lecture notes are available at `stat.snu.ac.kr/theostat`. Also I referred to following books when I write this note. The list would be updated continuously.

- *Probability: Theory and Examples, R.Durrett*
- *Mathematical Statistics (in Korean), W.C.Kim*

If you want to correct typo or mistakes, please contact to: `joonpyokim@snu.ac.kr`

Chapter 1

Asymptotic Approximations

1.1 Consistency

1.1.1 Preliminary for the chapter

Definition 1.1.1 (Notations)**.** *Let Θ be a parameter space. Then we consider a 'random variable' X on the probability space $(\Omega, \mathcal{F}, P_\theta)$ which is a function*

$$X : (\Omega, \mathcal{F}, P_\theta) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_\theta^X),$$

where $P_\theta^X := P_\theta \circ X^{-1}$. Note that P_θ is a probability measure from the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$. For the convenience, now we omit the explanation of fundamental setting.

Definition 1.1.2 (Convergence)**.** *Let $\{X_n\}$ be a sequence of random variables.*

1. $X_n \xrightarrow[n\to\infty]{a.s} X$ *if* $P\left(\lim_{n\to\infty} X_n = X\right) = 1 \Leftrightarrow P(|X_n - X| > \epsilon \text{ } i.o.) = 0 \ \forall \epsilon > 0$
$$\Leftrightarrow \lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty} (|X_n - X| > \epsilon)\right) = 0 \ \forall \epsilon > 0$$

2. $X_n \xrightarrow[n\to\infty]{P} X$ *if* $\forall \epsilon > 0 \ P(|X_n - X| > \epsilon) \to 0$ *as* $n \to \infty$.

Proposition 1.1.3. $X_n \xrightarrow[n\to\infty]{P} X$ *if and only if for any subsequence $\{n_k\} \subseteq \{n\}$ there is a further subsequence $\{n_{k_j}\} \subseteq \{n_k\}$ such that $X_{n_{k_j}} \xrightarrow[j\to\infty]{a.s.} X$.*

Proof. Durrett, p.65. □

Definition 1.1.4 (Consistency)**.** *$\hat{q}_n = q_n(X_1, \cdots, X_n)$ is consistent estimator of $q(\theta)$ if*

$$\hat{q}_n \xrightarrow[n\to\infty]{P_\theta} q(\theta)$$

for any $\theta \in \Theta$. (We don't know what is the true parameter.)

Remark 1.1.5. There are some tools to obtain consistency.

1. $Var(Z_n) \to 0$, $EZ_n \to \mu$ as $n \to \infty \Rightarrow Z_n \xrightarrow[n\to\infty]{P} \mu$.

$$\begin{aligned} \because P(|Z_n - \mu| > \epsilon) &\le P(|Z_n - EZ_n| + |EZ_n - \mu| > \epsilon) \\ &\le P(|Z_n - EZ_n| > \epsilon/2) + \underbrace{P(|EZ_n - \mu| > \epsilon/2)}_{=0 \text{ for sufficiently large } n} \\ &\le \frac{4}{\epsilon^2} Var(Z_n) \to 0 \end{aligned}$$

2. WLLN: $X_1, \cdots, X_n$: i.i.d. and $E|X_1| < \infty \Rightarrow \overline{X}_n \xrightarrow[n\to\infty]{P} EX_1$.

3. If $Z_n \xrightarrow[n\to\infty]{P} Z$ and g is continuous on the support of Z, then $g(Z_n) \xrightarrow[n\to\infty]{P} g(Z)$. Note that uniform convergence of g implies this directly, and for the general case, we can use Proposition 1.1.3.

4. Followings are the corollary of 3. Or, we can prove it directly. Suppose that $Y_n \xrightarrow[n\to\infty]{P} Y$ and $Z_n \xrightarrow[n\to\infty]{P} Z$. Then,

   (a) $Y_n + Z_n \xrightarrow[n\to\infty]{P} Y + Z$.

   (b) $Y_n Z_n \xrightarrow[n\to\infty]{P} YZ$.

   (c) $Y_n/Z_n \xrightarrow[n\to\infty]{P} Y/Z$ provided that $Z \neq 0$ P-a.s..

   Proof. (b) Note that $|Y_nZ_n - YZ| \le |Y_n||Z_n - Z| + |Z||Y_n - Y| \le |Y_n - Y||Z_n - Z| + |Y||Z_n - Z| + |Z||Y_n - Y|$. Now for any $\eta > 0$ there exists $M > 0$ such that $P(|Y| > M) \le \eta$ and $P(|Z| > M) \le \eta$. Now,

$$\begin{aligned} P(|Y_nZ_n - YZ| > \epsilon) &\le P(|Y_n||Z_n - Z| > \epsilon/2) + P(|Z||Y_n - Y| > \epsilon/2) \\ &\le P(|Y_n - Y||Z_n - Z| > \epsilon/4) + P(|Y||Z_n - Z| > \epsilon/4) + P(|Z||Y_n - Y| > \epsilon/2) \end{aligned}$$

   and note that $P(|Y||Z_n - Z| > \epsilon/4) = P(|Y||Z_n - Z| > \epsilon/4, |Y| > M) + P(|Y||Z_n - Z| > \epsilon/4, |Y| \le M) \le \eta + P(|Z_n - Z| \ge \epsilon/4M)$. Thus

$$\limsup_{n\to\infty} P(|Y||Z_n - Z| > \epsilon/4) \le \eta$$

and similarly

$$\limsup_{n\to\infty} P(|Z||Y_n - Y| > \epsilon/2) \le \eta.$$

Now, since

$$\begin{aligned}P(|Y_n - Y||Z_n - Z| > \epsilon/4) &= P(|Y_n - Y||Z_n - Z| > \epsilon/4, |Y_n - Y| > \sqrt{\epsilon/4})\\ &\quad + P(|Y_n - Y||Z_n - Z| > \epsilon/4, |Y_n - Y| \le \sqrt{\epsilon/4})\\ &\le P(|Y_n - Y| > \sqrt{\epsilon/4}) + P(|Z_n - Z| \ge \sqrt{\epsilon/4}) \to 0\end{aligned}$$

as $n \to \infty$, we get

$$\limsup_{n\to\infty} P(|Y_n Z_n - YZ| > \epsilon) \le 2\eta.$$

Finally, since $\eta > 0$ was arbitrary, we get the result.
(c) By (b), it's sufficient to show that $Z_n^{-1} \xrightarrow[n\to\infty]{P} Z^{-1}$. Since $P(Z = 0) = 0$, for any $\eta > 0$ there exists $\delta > 0$ such that $P(|Z| \le \delta) \le \eta$. (If not, $\exists \eta > 0$ such that $\forall \delta > 0$ $P(|Z| \le \delta) > \eta$. Then by continuity of measure, $P(\bigcup_{\delta>0}(|Z| \le \delta)) = P(Z = 0) \ge \eta > 0$. Contradiction.) Thus

$$\begin{aligned}P\left(\left|\frac{1}{Z_n} - \frac{1}{Z}\right| > \epsilon\right) &= P\left(\frac{|Z_n - Z|}{|Z_n Z|} > \epsilon\right)\\ &\le P\left(\frac{|Z_n - Z|}{|Z|(|Z| - |Z_n - Z|)} > \epsilon\right)\\ &\le \underbrace{P\left(\frac{|Z_n - Z|}{|Z|(|Z| - |Z_n - Z|)} > \epsilon, |Z| > \delta, |Z_n - Z| \le \delta/2\right)}_{\le P(|Z_n - Z| > \frac{\delta^2}{2}\epsilon) \xrightarrow[n\to\infty]{} 0}\\ &\quad + \underbrace{P(|Z| \le \delta)}_{\le \eta} + \underbrace{P(|Z_n - Z| > \delta/2)}_{\xrightarrow[n\to\infty]{} 0}\end{aligned}$$

and hence

$$\limsup_{n\to\infty} P\left(\left|\frac{1}{Z_n} - \frac{1}{Z}\right| > \epsilon\right) \le \eta$$

holds. Note that $\eta > 0$ was arbitrary. □

Definition 1.1.6 (Probabilistic O-notation)**.** *Let X_n be a sequence of r.v.'s.*

1. *$X_n = O_p(1)$ if $\lim_{c\to\infty} \sup_n P(|X_n| > c) = 0 \Leftrightarrow \lim_{c\to\infty} \limsup_{n\to\infty} P(|X_n| > c) = 0$. ("Bounded in probability")*

2. $X_n = o_p(1)$ *if* $X_n \xrightarrow[n\to\infty]{P} 0$.

3. $X_n = O_p(a_n)$ *if* $X_n/a_n = O_p(1)$*, and* $X_n = o_p(a_n)$ *if* $X_n/a_n = o_p(1)$.

Proposition 1.1.7. *If* $X_n \xrightarrow[n\to\infty]{d} X$ *for some* X*, then* $X_n = O_p(1)$.

Proof. For given $\epsilon > 0$, there exists c such that $P(|X| > c) < \epsilon/2$. For such c, $P(|X_n| > c) \to P(|X| > c)$, so $\exists N$ s.t.

$$\sup_{n>N} |P(|X_n| > c) - P(|X| > c)| < \frac{\epsilon}{2}.$$

Thus $\sup_{n>N} P(|X_n| > c) < \epsilon$. For $n = 1, 2, \cdots, N$, there exists c_n such that $P(|X_n| > c_n) < \epsilon$, and letting $c^* = \max(c_1, \cdots, c_N, c)$, we get $\sup_n P(|X_n| > c^*) < \epsilon$. □

Example 1.1.8 (Simple Linear Regression)**.** Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} (0, \sigma^2)$. Also assume that $x_1, \cdots, x_n$ are known and not all equal. Note that

$$\hat\beta_{1,n} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Since $E(\hat\beta_{1,n}) = \beta_1$ and $Var(\hat\beta_{1,n}) = \sigma^2/S_{xx}$, we obtain consistency

$$\hat\beta_{1,n} \xrightarrow[n\to\infty]{P_{\beta,\sigma^2}} \beta_1$$

provided that $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2 \to \infty$ as $n \to \infty$.

Example 1.1.9 (Sample correlation coefficient)**.** Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be random sample from the population

$$EX_1 = \mu_1,\ EY_1 = \mu_2,\ Var(X_1) = \sigma_1^2 > 0,\ Var(Y_1) = \sigma_2^2 > 0,\ \text{and } Corr(X_1, Y_1) = \rho.$$

Then by WLLN we get

$$(\overline{X}, \overline{Y}, \overline{X^2}, \overline{Y^2}, \overline{XY}) \xrightarrow[n\to\infty]{P} (EX_1, EY_1, EX_1^2, EY_1^2, EX_1Y_1).$$

Since the function

$$g(u_1, u_2, u_3, u_4, u_5) = \frac{u_5 - u_1u_2}{\sqrt{u_3 - u_1^2}\sqrt{u_4 - u_2^2}}$$

is continuous *at* $(EX_1, EY_1, EX_1^2, EY_1^2, EX_1Y_1)$, we get

$$\hat{\rho}_n = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{\overline{X^2} - \overline{X}^2}\sqrt{\overline{Y^2} - \overline{Y}^2}} \xrightarrow[n\to\infty]{P} \rho.$$

Remark 1.1.10. Note that, if $X_n \xrightarrow[n\to\infty]{P} c$ where c is a constant, then continuity of $g(x)$ at $x = c$ is sufficient for consistency $g(X_n) \xrightarrow[n\to\infty]{P} g(c)$. It is trivial from the definition of continuity.

Example 1.1.11. Let $X_1, \cdots, X_n$ be a random sample from a population with cdf F. Then we use an *empirical distribution function*

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$$

for estimation of F. Then by WLLN, for each x, $\hat{F}_n(x)$ is consistent estimator for $F(x)$,

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{P} F(x).$$

Remark 1.1.12. Note that in this case, we can say more strong result, which is known as *Glivenko-Cantelli theorem*:

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow[n\to\infty]{P} 0.$$

Sketch of proof is given here. Since $\hat{F}_n$ and F are nondecreasing and bounded, we can partition $[0, 1]$, which is a range of both functions, into finite number of intervals, and then each interval has a well-defined inverse image which is an interval. For whole proof, see Durrett, p.76.

1.1.2 FSE and MLE in Exponential Families

FSE

Recall that FSE of $\nu(F)$ is defined as $\nu(\hat{F}_n)$. One example of FSE is MME: to estimate $EX^j =: \nu_j(F) =: \int x^j dF(x)$, we use

$$\hat{m}_j = \nu_j(\hat{F}_n) = \int x^j d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i^j.$$

By WLLN we have $(\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_k)^T \xrightarrow[n\to\infty]{P} (m_1, m_2, \cdots, m_k)^T$ where $m_j = EX^j$, so we can obtain consistency of MME easily.

Proposition 1.1.13. *Let $q = h(m_1, m_2, \cdots, m_k)$ be a parameter of interest where m_j's are* *population moments. Then for MME*

$$\hat{q}_n = h(\hat{m}_1, \cdots, \hat{m}_k)$$

based on a random sample $X_1, \cdots, X_n$*,*

$$\hat{q}_n \xrightarrow[n\to\infty]{P} q$$

holds, provided that h *is continuous at* $(m_1, \cdots, m_k)^T$*.*

We can do similar work in FSE $\nu(F)$. Note that in here, ν is a functional, so we may define a continuity of functional. We may use sup norm as a metric in the space of distribution functions.

Definition 1.1.14. *Let* $\mathcal{F}$ *be a space of distribution functions. In this space, we give the norm* $\|\cdot\|$ *as a sup norm*

$$\|F\| = \sup_x |F(x)|.$$

Then metric is given as

$$\|F - G\| = \sup_x |F(x) - G(x)|.$$

Also, we say that a functional $\nu : \mathcal{F} \to \mathbb{R}$ *is continuous at* F *if for any* $\epsilon > 0$ *there exists* $\delta > 0$ *such that*

$$\|G - F\| < \delta \Rightarrow |\nu(G) - \nu(F)| < \epsilon.$$

Remark 1.1.15. Note that since $\|\hat{F}_n - F\| \to 0$ as $n \to \infty$ from Glivenko-Cantelli theorem, we get consistency of FSE

$$\nu(\hat{F}_n) \xrightarrow[n\to\infty]{P} \nu(F)$$

provided that ν is continuous at F. In many cases, showing continuity may be difficult problem.

Example 1.1.16 (Best Linear Predictor)**.** Let $X_1, \cdots, X_n$ be k-dimensional i.i.d. r.v.'s, and $Y_1, \cdots, Y_n$ be i.i.d. 1-dim random variable. Then we know that

$$BLP(x) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu_1),$$

where

$$E\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } Var\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Thus for sample variance

$$S_{11} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

$$S_{12} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})^T = S_{21}^T$$

$$S_{22} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

we obtain FSE for BLP,

$$\widehat{BLP}^{FSE}(x) = \overline{Y} + S_{21}S_{11}^{-1}(x - \overline{X}).$$

Note that it is same as sample linear regression line. Detail is given in next proposition.

Proposition 1.1.17.

(a) Solution of minimizing problem

$$(\beta_0^*, \boldsymbol{\beta}_1^*)^T = \underset{\beta_0, \boldsymbol{\beta}_1}{\arg\min} E(Y - \beta_0 - \boldsymbol{\beta}_1^T X)^2$$

is

$$BLP(x) := \beta_0^* + \boldsymbol{\beta}_1^{*T} x = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu_1).$$

(b) For $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ and design matrix $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1)$ where $\boldsymbol{X}_1 = (X_1, \cdots, X_n)^T$, LSE is

$$\hat{\boldsymbol{\beta}}_1 = S_{11}^{-1}S_{12} \text{ and } \hat{\beta}_0 = \overline{Y} - \overline{X}^T\hat{\boldsymbol{\beta}}_1.$$

Proof. (a) Two approaches are given. First one is direct proof: It is clear because of

$$\begin{aligned} E(Y - \beta_0 - \boldsymbol{\beta}_1^T X)^2 &= E[(Y - \mu_2) - \boldsymbol{\beta}_1^T(X - \mu_1)]^2 + [\mu_2 - \beta_0 - \boldsymbol{\beta}_1^T\mu_1]^2 \\ &= \Sigma_{22} - 2\boldsymbol{\beta}_1^T\Sigma_{12} + \boldsymbol{\beta}_1^T\Sigma_{11}\boldsymbol{\beta}_1 + [\beta_0 - (\mu_2 - \boldsymbol{\beta}_1^T\mu_1)]^2. \end{aligned}$$

Second approach uses projection in $\mathcal{L}^2$ space. For convenience, suppose $EX = 0$ and $EY = 0$. Then $(\beta_0^*, \boldsymbol{\beta}_1^*)^T$ should satisfy

$$\langle \beta_0 + \boldsymbol{\beta}_1^T X, Y - \beta_0^* - \boldsymbol{\beta}_1^{*T} X\rangle = 0 \ \forall \beta_0, \beta_1.$$

It yields that

$$\beta_0^* = 0, \ \boldsymbol{\beta}_1^* = \left(E(XX^T)\right)^{-1} E(XY).$$

(b) $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1$ should satisfy $\mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y}|\mathcal{C}(\boldsymbol{X}))$. For $\mathcal{X}_1 = \boldsymbol{X}_1 - \Pi(\boldsymbol{X}_1|\mathcal{C}(\mathbf{1})) = \boldsymbol{X}_1 - \mathbf{1}\overline{X}^T$,

$$\mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \mathbf{1}\left(\hat{\beta}_0 + \frac{\mathbf{1}^T\boldsymbol{X}_1}{n}\hat{\boldsymbol{\beta}}_1\right) + \mathcal{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y}|\mathcal{C}(\mathbf{1})) + \Pi(\boldsymbol{Y}|\mathcal{C}(\mathbf{X_1}))$$

we get

$$\hat{\beta}_0 = \overline{Y} - \overline{X}^T\hat{\boldsymbol{\beta}}_1 \text{ and } \hat{\boldsymbol{\beta}}_1 = (\mathcal{X}_1^T\mathcal{X}_1)^{-1}\mathcal{X}_1^T\boldsymbol{Y}.$$

Now $\mathcal{X}_1^T\mathcal{X}_1 = S_{11}$ and $\mathcal{X}_1^T\boldsymbol{Y} = S_{12}$ ends the proof. □

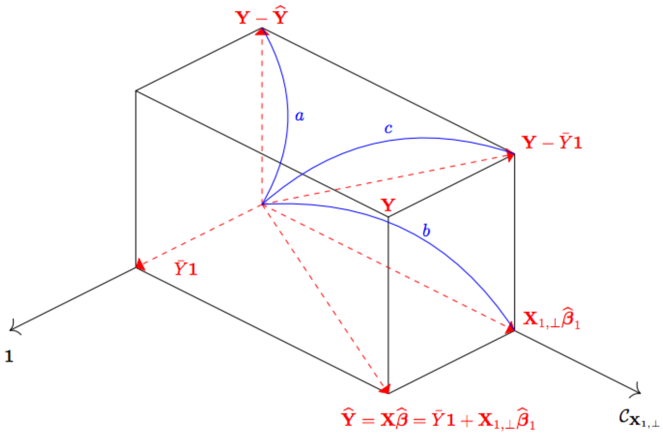

Figure 1.1: Regression with intercept. Image from Lecture Note.

Example 1.1.18 (Multiple Correlation Coefficient)**.** We define a *multiple correlation coefficient (MCC)* as

$$\rho = \max_{\beta_0, \boldsymbol{\beta}_1} \mathrm{Corr}(Y, \beta_0 + \boldsymbol{\beta}_1^T X)$$

and sample MCC is

$$\hat{\rho}_n = \max_{\beta_0, \boldsymbol{\beta}_1} \widehat{\mathrm{Corr}}(Y, \beta_0 + \boldsymbol{\beta}_1^T X).$$

Note that,

$$\begin{aligned}
\mathrm{Corr}(Y, \beta_0 + \boldsymbol{\beta}_1^T X) &= \mathrm{Corr}(Y - \mu_2, \boldsymbol{\beta}_1^T(X - \mu_1)) \\
&= \frac{\Sigma_{21}\boldsymbol{\beta}_1}{\sqrt{\Sigma_{22}}\sqrt{\boldsymbol{\beta}_1^T\Sigma_{11}\boldsymbol{\beta}_1}} \\
&= \frac{(\Sigma_{11}^{-1/2}\Sigma_{12})^T(\Sigma_{11}^{1/2}\boldsymbol{\beta}_1)}{\sqrt{\Sigma_{22}}\sqrt{\boldsymbol{\beta}_1^T\Sigma_{11}\boldsymbol{\beta}_1}}
\end{aligned}$$

$$\leq \sqrt{\frac{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}{\Sigma_{22}}}$$

holds by Cauchy-Schwarz inequality, and equality holds when $\boldsymbol{\beta}_1 = \Sigma_{11}^{-1}\Sigma_{12}$. Thus population MCC is obtained as

$$\rho = \sqrt{\frac{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}{\Sigma_{22}}}.$$

Meanwhile, sample correlation is obtained as

$$\widehat{\text{Corr}}(\boldsymbol{Y}, \beta_0 + \boldsymbol{\beta}_1^T\boldsymbol{X}) = \frac{\langle \boldsymbol{Y} - \overline{Y}\mathbf{1}, (\boldsymbol{X} - \mathbf{1}\overline{X}^T)\boldsymbol{\beta}_1\rangle}{\|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|\|(\boldsymbol{X} - \mathbf{1}\overline{X}^T)\boldsymbol{\beta}_1\|}$$

so it is the cosine of the angle between the two rays, $\boldsymbol{Y} - \overline{Y}\mathbf{1}$ and $\mathcal{X}_1\boldsymbol{\beta}_1$. Its maximal value is attaiend by $\mathcal{X}_1\hat{\boldsymbol{\beta}}_1 = \Pi(\boldsymbol{Y} - \overline{Y}\mathbf{1}|\mathcal{C}(\mathcal{X}_1))$. Thus,

$$\hat{\rho}^2 = \frac{SSR}{SST} = \frac{\hat{\boldsymbol{\beta}}_1^T\mathcal{X}_1^T\mathcal{X}_1\hat{\boldsymbol{\beta}}_1}{\|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|^2} = \frac{S_{21}S_{11}^{-1}S_{12}}{S_{22}}.$$

Example 1.1.19 (Sample Proportions)**.** Let $(X_1, \cdots, X_k)^T \sim Multi(n, p)$, where $p \in \Theta := \{(p_1, \cdots, p_k)^T : \sum_{i=1}^k p_i = 1,\ p_i \geq 0\ (i = 1, 2, \cdots, k)\}$. We estimate p with sample proportion

$$\hat{p}_n = \left(\frac{X_1}{n}, \cdots, \frac{X_k}{n}\right)^T.$$

Then,

(a) $\hat{p}_n$ is consistent estimator of p, i.e.,

$$\forall\epsilon > 0,\ \sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \epsilon) \xrightarrow[n\to\infty]{} 0.$$

(b) $q(\hat{p}_n)$ is consistent estimator of $q(p)$ provided that q is (uniformly) continuous on Θ.

Proof. (a) Note that there exists a constant $C > 0$ such that

$$\begin{aligned}
\sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \epsilon) &\leq \sup_{p\in\Theta} \frac{E|\hat{p}_n - p|^2}{\epsilon^2} \\
&= \sup_{p\in\Theta} \sum_{i=1}^k \frac{p_i(1-p_i)}{n\epsilon^2} \\
&\leq \frac{C}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

so we get the desired result. Note that first inequality is from Chebyshev's inequality.

(b) Note that q is uniformly continuous on Θ, since Θ is closed and bounded. Thus the assertion holds. More precisely, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|p' - p| < \delta,\ p, p' \in \Theta \Rightarrow |q(p') - q(p)| < \epsilon.$$

Therefore, we get

$$\sup_{p\in\Theta} P_p(|q(\hat{p}_n) - q(p)| \geq \epsilon) \leq \sup_{p\in\Theta} P_p(|\hat{p}_n - p| \geq \delta) \xrightarrow[n\to\infty]{} 0.$$

$\square$

MLE in exponential families

Consider a random variable X with pdf in canonical exponential family

$$q_\eta(x) = h(x)\exp(\eta^T T(x) - A(\eta))I_{\mathcal{X}}(x),\ \eta \in \mathcal{E},$$

where $\mathcal{E}$ is natural parameter space in $\mathbb{R}^k$. Our goal is to show consistency of MLE in canonical exponential family.

Theorem 1.1.20. *Let*

$$q_\eta(x) = h(x)\exp(\eta^T T(x) - A(\eta))I_{\mathcal{X}}(x),\ \eta \in \mathcal{E}$$

be a canonical exponential family with natural parameter space $\mathcal{E} \subseteq \mathbb{R}^k$. Further assume

(i) $\mathcal{E}$ is open.

(ii) The family is of rank k.

(iii) $t_0 := T(x) \in C^0$, where C denotes the smallest convex set containing the support of $T(X)$, and C^0 be its interior.

Then the unique ML estimate $\hat{\eta}(x)$ exists and is the solution of the likelihood equation

$$\dot{l}_x(\eta) = T(x) - \dot{A}(\eta) = 0.$$

Remark 1.1.21. Note that in (iii), x is the observation of X, so t_0 is the observation of $T(X)$. It is reasonable to consider t_0 because ML estimate only depends on t_0. Also, recall that (ii) means

$$\begin{aligned} &\nexists a \neq 0 \ s.t. \ \left[P_\eta(a^T(T(x)-\mu)=0)=1 \text{ for some } \eta \in \mathcal{E}\right] \\ \Leftrightarrow &\nexists a \neq 0 \ s.t. \ \left[Var_\eta(a^T T(x))=0 \text{ for some } \eta \in \mathcal{E}\right] \\ \Leftrightarrow &\ddot{A}(\eta) \text{ is positive definite } \forall \eta \in \mathcal{E}. \end{aligned}$$

To prove this, we need some preparation.

Lemma 1.1.22.

(a) ("Supporting Hyperplane Theorem") Let $C \subseteq \mathbb{R}^k$ be a convex set, and C^0 be its interior. Then for $t_0 \notin C$ or $t_0 \in \partial C$,

$$\exists a \neq 0 \ s.t. \ [a^T t \geq a^T t_0 \ \forall t \in C].$$

Conversely, for $t_0 \in C^0$,

$$\nexists a \neq 0 \ s.t. \ [a^T t \geq a^T t_0 \ \forall t \in C].$$

(b) Let $P(T \in \mathcal{T}) = 1$ and $E(\max_i |T_i|) < \infty$. (i.e., $\mathcal{T}$ is support of T.) Then for a convex hull C of $\mathcal{T}$, we get $ET \in C^0$.

(c) Assume the above exponential family model with open $\mathcal{E}$. Then the ML estimate exists if the log-likelihood approaches $-\infty$ on the boundary.

Proof. (a) Only second part will be given. (For the first part, see supplementary note.) Let $t_0 \in C^0$. Then $\exists \delta > 0$ such that $B(t_0, \delta) \subseteq C^0$, since C^0 is open. Note that for any u s.t. $\|u\| = 1$, we get

$$t_0 - \frac{\delta}{2}u, \ t_0 + \frac{\delta}{2}u \in B(t_0, \delta) \subseteq C.$$

If $\exists a \neq 0$ such that $a^T t \geq a^T t_0 \ \forall t \in C$, then

$$a^T\left(t_0 - \frac{\delta}{2}u\right) \geq a^T t_0, \ a^T\left(t_0 + \frac{\delta}{2}u\right) \geq a^T t_0$$

holds for $u = a/|a|$, which yields contradiction. (Note that convexity condition is not used)

(b) Note that since C is a convex set, $\mu := ET \in C$ holds. (Convex set contains average of itself) Assume $\mu \notin C^0$. Then $\mu \in \partial C$. Then by (a), $\exists a \neq 0$ such that $a^T t \geq a^T \mu$ for any $t \in C$.

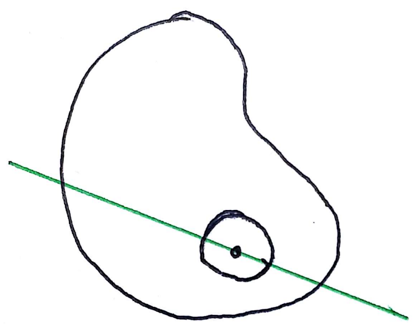

Figure 1.2: Proof of (a)

It implies that, $\exists a \neq 0$ such that $P(a^T(T-\mu)\geq 0)=1$, since $\mathcal{T}\subseteq C$. It implies that

$$P(a^T(T-\mu)=0)=1,$$

by the fact that

$$f\geq 0,\ \int f d\mu = 0 \Rightarrow f=0\ \mu-a.e..$$

It is contradictory to (ii), which is full rank condition of the exponential family.

(c) Done in TheoStat I.

Proof of theorem. By lemma, it's sufficient to show that:

(1) $l(\theta)$ diverges to $-\infty$ at the boundary. (Existence)

(2) Uniqueness

Note that Uniqueness is clear since $l_x(\eta)$ is $\mathcal{C}^2$ function and strictly concave from $\ddot{A}(\eta)>0$. Thus, our claim is

Claim. $l(\theta)$ approches $-\infty$ on the boundary $\partial\mathcal{E}$.

Let $\eta^0\in\partial\mathcal{E}$. Then there is $\eta_n \xrightarrow[n\to\infty]{} \eta^0$ such that $\eta_n\in\mathcal{E}$. Now our claim is, for any such sequence η_n, we get $l_x(\eta_n)\xrightarrow[n\to\infty]{} -\infty$. Note that $|\eta_n|\xrightarrow[n\to\infty]{}\infty$ or $\sup|\eta_n|<\infty$. Also note that, for both cases, from $l_x(\eta)=\log h(x)+\eta^T T(x)-A(\eta)$ and $e^{A(\eta)}=\int_{\mathcal{X}} h(x)e^{\eta^T T(x)}d\mu(x)$, we get

$$-l_x(\eta_n)+\log h(x) = A(\eta_n)-\eta_n^T t_0$$

$$= \log \int_{\mathcal{X}} \exp\left(\eta_n^T (T(y) - t_0)\right) h(y) d\mu(y).$$

<u>Case 1.</u> $|\eta_n| \to \infty$.

Then since

$$\begin{aligned}
\int_{\mathcal{X}} e^{\eta_n^T (T(y)-t_0)} h(y) d\mu(y) &\ge \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} e^{|\eta_n| \cdot \frac{\eta_n^T}{|\eta_n|}(T(y)-t_0)} h(y) d\mu(y) \\
&\ge \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} e^{|\eta_n|/k} h(y) d\mu(y) \\
&= \exp(|\eta_n|/k) \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y),
\end{aligned}$$

if we can conclude

$$\inf_n \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y) > 0,$$

by the assumption $|\eta_n| \to \infty$, we get $l_x(\eta_n) \to -\infty$. Note that if

$$\inf_{u: \|u\|=1} \int_{u^T (T(y)-t_0) > 0} h(y) d\mu(y) > 0,$$

then

$$\inf_n \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > 0} h(y) d\mu(y) > 0,$$

and from

$$\inf_n \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y) \xrightarrow[k\to\infty]{} \inf_n \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > 0} h(y) d\mu(y),$$

we get $\exists \epsilon > 0$ & k s.t.

$$\inf_n \int_{\frac{\eta_n^T}{|\eta_n|}(T(y)-t_0) > \frac{1}{k}} h(y) d\mu(y) > \epsilon$$

and the assertion holds. So our claim is:

<u>**Claim.**</u> $\inf_{u: \|u\|=1} \int_{u^T (T(y)-t_0) > 0} h(y) d\mu(y) > 0.$

Assume not. If

$$\inf_{u: \|u\|=1} \int_{u^T (T(y)-t_0) > 0} h(y) d\mu(y) = 0,$$

then since $\{u : \|u\| = 1\}$ is compact, there exists $u_0 \in \{u : \|u\| = 1\}$ such that

$$\int_{u_0^T(T(y)-t_0)>0} h(y)d\mu(y) = 0.$$

It implies $h(y) = 0$ on the set $\{y : u_0^T(T(y) - t_0) > 0\}$ μ-a.e., and hence

$$\int_{u_0^T(T(y)-t_0)>0} h(y)e^{\eta^T T(y)-A(\eta)}d\mu(y) = 0,$$

which implies that

$$P_\eta(u_0^T(T(X) - t_0) > 0) = 0.$$

Thus, we get

$$P_\eta(u_0^T(T(X) - t_0) \le 0) = 1,$$

which is equivalent to

$$u_0^T(t - t_0) \le 0 \ \forall t \in \mathcal{T}.$$

Since C is convex hull of $\mathcal{T}$, it implies

$$u_0^T(t - t_0) \le 0 \ \forall t \in C,$$

however, this yields contradiction to

$$\nexists a \neq 0 \ s.t. \ a^T(t - t_0) \le 0 \ \forall t \in C,$$

from $t_0 \in C^0$.

<u>Case 2.</u> $\sup|\eta_n| < \infty$

In this case, we get

$$\liminf_{n\to\infty} \int_{\mathcal{X}} e^{\eta_n^T(T(y)-t_0)}h(y)d\mu(y) \ge \int_{\mathcal{X}} e^{\eta^{0T}(T(y)-t_0)}h(y)d\mu(y) \stackrel{(*)}{=} \infty$$

by Fatou's lemma. $(*)$ holds because $\mathcal{E}$ is natural parameter space, and $\eta^0 \in \partial\mathcal{E}$ implies $\eta^0 \notin \mathcal{E}$, since $\mathcal{E}$ is open. Thus $-l_x(\eta_n) \xrightarrow[n\to\infty]{} \infty$.

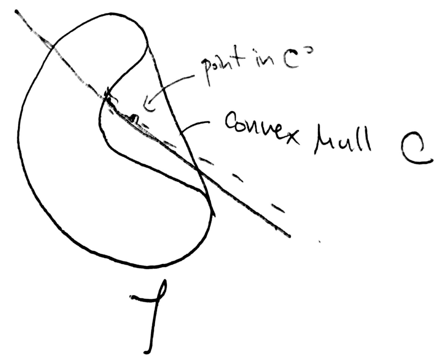

Figure 1.3: Convex hull of $\mathcal{T}$

□

Now we are ready to prove consistency.

Theorem 1.1.23. *Let $X_1, \cdots, X_n$ be a random sample from a population with pdf*

$$p_\eta(x) = h(x)\exp\{\eta^T T(x) - A(\eta)\} I_{\mathcal{X}}(x),\ \eta \in \mathcal{E}$$

where $\mathcal{E}$ is the natural parameter space in $\mathbb{R}^k$. Further, assume that

(i) $\mathcal{E}$ is open.

(ii) The family is of rank k.

Then, the followings hold:

(a) $P_\eta\left(\hat{\eta}_n^{MLE} \text{ exists}\right) \xrightarrow[n\to\infty]{} 1$

(b) $\hat{\eta}_n^{MLE}$ is consistent.

Proof. (a) Let $\overline{T}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^n T(X_i)$. Then by WLLN, we get

$$\lim_{n\to\infty} P_\eta(|\overline{T}_n - E_\eta T(X_1)| < \epsilon) = 1\ \forall \epsilon > 0.$$

Also note that $E_\eta T(X_1) \in C^0$, where C^0 is the interior of the convex hull of the support of $T(X_1)$. Then since C^0 is open, open ball $(|\overline{T}_n - E_\eta T(X_1)| < \epsilon)$ is contained in C^0 for sufficiently small $\epsilon > 0$, which implies

$$\lim_{n\to\infty} P_\eta(\overline{T}_n \in C^0) = 1.$$

Now consider $\overline{T}_n$ instead of $T(X_1)$ in previous theorems, and note that (convex hull of support of $\overline{T}_n$) = (convex hull of support of $T(X_1)$). Then we can find that

$$(\overline{T}_n \in C^0) \subseteq (\hat{\eta}_n^{MLE} \text{ exists})$$

and therefore

$$\lim_{n\to\infty} P_\eta(\hat{\eta}_n^{MLE} \text{ exists}) = 1.$$

(b) From $\ddot{A} > 0$, we get $\dot{A}(\eta)$ is one-to-one and continuous for any η. Then we get

$$(\overline{T}_n \in C^0) \subseteq (\hat{\eta}_n^{MLE} \text{ exists}) = (\dot{A}(\hat{\eta}_n^{MLE}) = \overline{T}_n)$$

and hence

$$\lim_{n\to\infty} P_\eta(\hat{\eta}_n^{MLE} = (\dot{A})^{-1}(\overline{T}_n)) = 1 \ \forall \eta \in \mathcal{E}. \tag{1.1}$$

Further, by inverse function theorem, and C^2 property of A, we have that $(\dot{A})^{-1}$ is continuous. Thus by WLLN and continuous mapping theorem,

$$(\dot{A})^{-1}(\overline{T}_n) \xrightarrow[n\to\infty]{P_\eta} (\dot{A})^{-1}(E_\eta T(X_1)) = (\dot{A})^{-1}(\dot{A}(\eta)) = \eta$$

and since $(\dot{A})^{-1}(\overline{T}_n) \approx \hat{\eta}_n^{MLE}$ in the sense of (1.1), we get

$$\lim_{n\to\infty} P_\eta(|\hat{\eta}_n^{MLE} - \eta| < \epsilon) = 1 \ \forall \epsilon > 0,$$

i.e., $\hat{\eta}_n^{MLE} \xrightarrow[n\to\infty]{P_\eta} \eta$. □

Now let's see some general results. Suppose we have $\lim_{n\to\infty} \Psi_n(\theta) = \Psi_0(\theta)$ and

$$\theta_n : \text{solution of } \Psi_n(\theta) = 0,\ \theta \in C\ (n = 1, 2, \cdots)$$
$$\theta_0 : \text{solution of } \Psi_0(\theta) = 0,\ \theta \in C.$$

Under what conditions, $\lim_{n\to\infty} \theta_n = \theta_0$? We need following four conditions:

Uniform convergence of Ψ_n, Continuity of Ψ_0, Uniqueness of θ_0, and Compactness of C.

Note that these are sufficient conditions *simultaneously*. Our goal is to obtain similar result for optimization.

Theorem 1.1.24. *Suppose that we have* $\lim_{n\to\infty} D_n(\theta) = D_0(\theta)$ *and*

$$\theta_n = \arg\min_{\theta\in C} D_n(\theta) \ (n = 1, 2, \cdots)$$

$$\theta_0 = \arg\min_{\theta\in C} D_0(\theta)$$

where D_n *and* D_0 *are deterministic functions. Also assume that*

(i) D_n *converges to* D_0 *uniformly.*

(ii) D_0 *is continuous on* C*.*

(iii) Minimizer θ_0 *is unique.*

(iv) C *is compact.*

Then $\lim_{n\to\infty} \theta_n = \theta_0$*.*

Proof. Assume not. In other words, $\theta_n \not\to \theta_0$. Then $\exists\epsilon > 0$ such that $|\theta_n - \theta_0| > \epsilon$ i.o.. It means that there is a subsequence $\{n'\} \subseteq \{n\}$ s.t. $|\theta_{n'} - \theta_0| > \epsilon \ \forall n'$. Now define

$$\Delta_n = \sup_{\theta\in C} |D_n(\theta) - D_0(\theta)|.$$

Then by **uniform convergence** of D_n, we get $\Delta_n \xrightarrow[n\to\infty]{} 0$. Now note that

$$\begin{aligned}
\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) &= \inf_{|\theta-\theta_0|>\epsilon} \{D_0(\theta) - D_{n'}(\theta) + D_{n'}(\theta)\} \\
&\le \inf_{|\theta-\theta_0|>\epsilon} \{|D_0(\theta) - D_{n'}(\theta)| + D_{n'}(\theta)\} \\
&\le \Delta_{n'} + \inf_{|\theta-\theta_0|>\epsilon} D_{n'}(\theta)
\end{aligned}$$

holds. Because minimization of $D_{n'}$ is achieved at $\theta_{n'} \in \{\theta : |\theta - \theta_0| > \epsilon\}$, we get

$$\begin{aligned}
\Delta_{n'} + \inf_{|\theta-\theta_0|>\epsilon} D_{n'}(\theta) &\le \Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} D_{n'}(\theta) \\
&\le \Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} \{|D_{n'}(\theta) - D_0(\theta)| + D_0(\theta)\} \\
&\le 2\Delta_{n'} + \inf_{|\theta-\theta_0|\le\epsilon} D_0(\theta) \\
&= 2\Delta_{n'} + D_0(\theta_0).
\end{aligned}$$

The last equality holds from $\theta_0 = \arg\min D_0(\theta)$ and $\theta_0 \in \{\theta : |\theta - \theta_0| \le \epsilon\}$. Thus

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) \le 2\Delta_{n'} + D_0(\theta_0)$$

holds, which implies

$$\frac{1}{2}\left(\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0)\right) \le \Delta_{n'}.$$

Letting $n' \to \infty$, we get

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0) = 0.$$

It is contradictory due to our claim that will be shown:

Claim. $\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0) > 0.$

Intuitively, since θ_0 is unique minimizer, our claim seems trivial, but we also need continuity and compactness condition to guarantee this. (For this see next remark.)

Note that, by definition of infimum, there is a sequence $\{\theta_k\} \subseteq \{\theta : |\theta - \theta_0| > \epsilon\} \cap C$ such that

$$\lim_{k\to\infty} D_0(\theta_k) = \inf_{|\theta-\theta_0|>\epsilon} D_0(\theta).$$

Now, by **compactness of C**, there is a subsequence $\{k'\} \subseteq \{k\}$ that makes $\theta_{k'}$ converge to some θ_0^* ("Bolzano-Weierstrass"), so with the abuse of notation, let $\theta_k \to \theta_0^*$ as $k \to \infty$. Then note that θ_0^* should belong to $\{\theta : |\theta - \theta_0| \ge \epsilon\} \cap C$, so $\theta_0^* \ne \theta_0$. Now, **continuity of D_0** makes

$$\lim_{k\to\infty} D_0(\theta_k) = D_0(\theta_0^*),$$

which implies

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) = D_0(\theta_0^*).$$

Therefore, by **uniqueness of minimizer**, $D_0(\theta_0^*) > D_0(\theta_0)$, and combining to above result we can obtain

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) > D_0(\theta_0).$$

□

Remark 1.1.25. See next figures. Each example tells that we need continuity and compactness, respectively.

Remark 1.1.26. For deterministic case, one can give an alternative proof. Suppose $\theta_n \not\to \theta_0$. Then since C is compact, we can find a subsequence $\{\theta_{nk}\}$ such that $\theta_{nk} \to \theta_0^*$, $\theta_0^* \ne \theta_0$. (If any

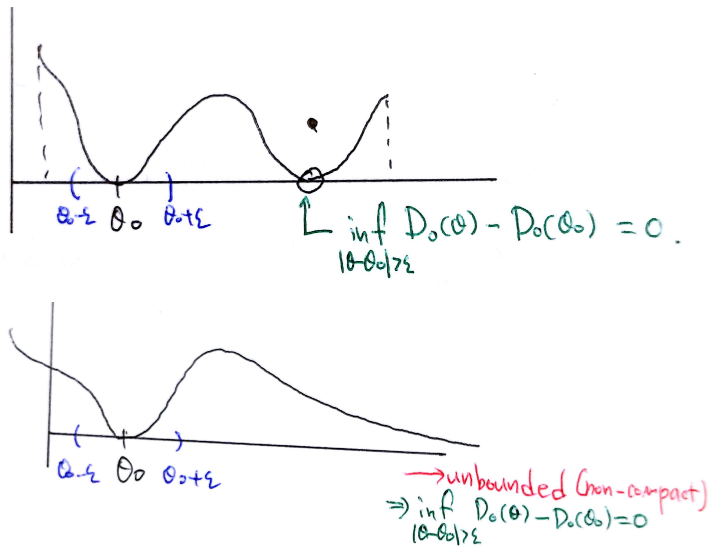

Figure 1.4: Continuity and Compactness are needed.

convergent subsequence converges to θ_0, then origin sequence should converge to θ_0.) Now for sufficiently large n_k,

$$\sup_{\theta\in C}|D_{nk}(\theta)-D_0(\theta)|<\frac{\epsilon}{3}$$

holds, so

$$\begin{aligned}
D_0(\theta_0) &\geq D_{nk}(\theta_0)-\frac{\epsilon}{3} \ (\because \text{uniform convergence})\\
&\geq D_{nk}(\theta_{nk})-\frac{\epsilon}{3} \ (\because \text{minimizer})\\
&\geq D_0(\theta_{nk})-\frac{2}{3}\epsilon \ (\because \text{uniform convergence})\\
&\geq D_0(\theta_0^*)-\epsilon \ (\because D_0(\theta_{nk})\to D_0(\theta_0^*) \text{ from continuity of } D_0)
\end{aligned}$$

and hence taking $\epsilon\searrow 0$ gives $D_0(\theta_0)\geq D_0(\theta_0^*)$, which is contradictory to uniqueness of θ_0.

In fact, our real goal was, to get the similar result for *random* D_n.

Theorem 1.1.27. *Let D_n be a sequence of random functions, and D_0 be deterministic. Similarly, define*

$$\begin{aligned}
\hat\theta_n &= \arg\min_{\theta\in C} D_n(\theta) \ (n=1,2,\cdots)\\
\theta_0 &= \arg\min_{\theta\in C} D_0(\theta).
\end{aligned}$$

Now suppose that

*(i) D_n converges in probability to D_0 **uniformly**. It means that,*

$$\sup_{\theta\in C}|D_n(\theta)-D_0(\theta)| \xrightarrow[n\to\infty]{P} 0.$$

(ii) D_0 is continuous on C.

(iii) Minimizer θ_0 is unique.

(iv) C is compact.

Then $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_0$.

Proof. Note that in the proof of theorem 1.1.24, we did not used convergence in deriving

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) \le 2\Delta_{n'} + D_0(\theta_0).$$

Rather, we only used $|\theta_{n'}-\theta_0|>\epsilon$. (Convergence is used when deriving $\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta)-D_0(\theta_0)$) Thus,

$$|\hat{\theta}_n-\theta_0|>\epsilon \Rightarrow \inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) \le 2\Delta_{n'}+D_0(\theta_0) \Rightarrow \Delta_n \ge \frac{1}{2}\left(\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta)-D_0(\theta_0)\right)$$

holds. Define

$$\frac{1}{2}\left(\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta)-D_0(\theta_0)\right) =: \delta(\epsilon).$$

Then, we get

$$\left(|\hat{\theta}_n-\theta_0|>\epsilon\right) \subseteq \left(\Delta_n \ge \delta(\epsilon)\right),$$

and therefore, by uniform P-convergence, $\Delta_n \xrightarrow[n\to\infty]{P} 0$ and hence

$$P(|\hat{\theta}_n-\theta_0|>\epsilon) \le P(\Delta_n \ge \delta(\epsilon)) \xrightarrow[n\to\infty]{} 0.$$

□

Example 1.1.28 (Consistency of MLE when Θ is finite)**.** Let $X_1,\cdots,X_n$ be a random sample from a population with pdf $f_\theta(\cdot)$, $\theta\in\Theta$. Assume that the parametrization is identifiable and $\Theta=\{\theta_1,\cdots,\theta_k\}$. Then

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0,$$

provided that

(0) (Identifiability) $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$

(1) (Kullback-Leibler divergence) $E_{\theta_0} \left| \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} \right| < \infty$.

Proof. Note that, we defined

$$\hat{\theta}_n^{MLE} = \arg\min_{\theta \in \Theta} D_n(\theta) \text{ for } D_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)},$$

and by Kullback-Leibler divergence,

$$\theta_0 = \arg\min_{\theta \in \Theta} D_0(\theta) \text{ for } D_0(\theta) = -E_{\theta_0} \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}.$$

Then,

(i) $\Theta = \{\theta_1, \cdots, \theta_k\}$ is compact.

(ii) θ_0 is unique minimizer of D_0. (For this, see next remark.)

(iii) Uniform convergence is achieved from

$$\begin{aligned} P_{\theta_0}\left\{ \max_{1 \le j \le k} |D_n(\theta_j) - D_0(\theta_j)| > \epsilon \right\} &= P_{\theta_0}\left\{ \bigcup_{1 \le j \le k} (|D_n(\theta_j) - D_0(\theta_j)| > \epsilon) \right\} \\ &\le \sum_{j=1}^{k} P_{\theta_0}\left(|D_n(\theta_j) - D_0(\theta_j)| > \epsilon\right) \\ &= o(1) \text{ by WLLN.} \end{aligned}$$

so we can derive the result similarly. In precise, it's sufficient to show

$$\inf_{|\theta-\theta_0|>\epsilon} D_0(\theta) - D_0(\theta_0) > 0$$

for ϵ s.t. $|\theta_n - \theta_0| > \epsilon$ i.o.. Uniqueness of θ_0 implies it clearly, because Θ is finite in here. Note that continuity of D_0 is not considered. □

Remark 1.1.29. *Kullback-Leibler divergence.* Since $1 + \log z \le z$, we get

$$\begin{aligned} -E_{\theta_0} \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} &= -\int \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} dP_{\theta_0} \\ &\ge 1 - \int_{S(\theta_0)} \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x) d\mu(x) \end{aligned}$$

$$\geq 0,$$

and hence $D_0(\theta) \geq 0$. In here $S(\theta_0) = \{x : f_{\theta_0}(x) > 0\}$ and $S(\theta) = \{x : f_\theta(x) > 0\}$. Note that $1 + \log z \leq z \Leftrightarrow z = 1$. Thus equality of $D_0(\theta) = 0$ holds if and only if

$$\begin{aligned} \frac{f_\theta(x)}{f_{\theta_0}(x)} &= 1 \ \mu - \text{a.e. on } S(\theta_0) \\ \text{and } \int_{S(\theta_0)} f_\theta(x) d\mu(x) &= 1. \end{aligned}$$

Since

$$\begin{aligned} 1 = \int_{S(\theta)} f_\theta(x) d\mu(x) &= \int_{S(\theta_0) \cup S(\theta)} f_\theta(x) d\mu(x) \\ &= \int_{S(\theta_0)} f_\theta(x) d\mu(x) + \int_{S(\theta) \setminus S(\theta_0)} f_\theta(x) d\mu(x) \end{aligned}$$

we get

$$\int_{S(\theta_0)} f_\theta(x) d\mu(x) = 1 \Leftrightarrow \int_{S(\theta) \setminus S(\theta_0)} f_\theta(x) d\mu(x) = 0.$$

However, by definition of the support, $f_\theta(x) > 0$ on $S(\theta) \setminus S(\theta_0)$, and hence

$$\int_{S(\theta) \setminus S(\theta_0)} f_\theta(x) d\mu(x) = 0 \Leftrightarrow \mu(S(\theta) \setminus S(\theta_0)) = 0.$$

Thus $D_0(\theta)$ holds if and only if

$$\begin{aligned} f_\theta(x) &= f_{\theta_0}(x) \ \mu - \text{a.e. on } S(\theta_0) \\ &\text{and } \mu(S(\theta) \setminus S(\theta_0)) = 0. \end{aligned}$$

However, note that

$$f_\theta(x) = f_{\theta_0}(x) \ \mu - \text{a.e. on } S(\theta_0) \text{ implies } \mu(S(\theta) \setminus S(\theta_0)) = 0,$$

because

$$\begin{aligned} 1 = \int_{S(\theta)} f_\theta(x) d\mu(x) &= \int_{S(\theta_0)} f_\theta(x) d\mu(x) + \int_{S(\theta) \setminus S(\theta_0)} f_\theta(x) d\mu(x) \\ &= \int_{S(\theta_0)} f_{\theta_0}(x) d\mu(x) + \int_{S(\theta) \setminus S(\theta_0)} f_\theta(x) d\mu(x) \end{aligned}$$

$$= 1 + \int_{S(\theta)\backslash S(\theta_0)} f_\theta(x) d\mu(x).$$

Therefore we get,

$$D_0(\theta) = 0 \Leftrightarrow f_\theta(x) = f_{\theta_0}(x)\ \mu - \text{a.e. on } S(\theta_0).$$

Now $\mu(S(\theta)\backslash S(\theta_0)) = 0$ implies $f_\theta(x) = f_{\theta_0}(x)\ \mu -$ a.e. on $S(\theta)\backslash S(\theta_0)$, and therefore $f_\theta(x) = f_{\theta_0}(x)\ \mu -$ a.e., if $f_\theta(x) = f_{\theta_0}(x)\ \mu -$ a.e. on $S(\theta_0)$.Therefore we get

$$D_0(\theta) = 0 \Leftrightarrow f_\theta(x) = f_{\theta_0}(x)\ \mu - \text{a.e.} \Leftrightarrow \theta = \theta_0\ (\because \text{identifiability}).$$

It means that θ_0 is unique minimizer of $D_0(\theta)$.

Example 1.1.30 (Consistency of MCE)**.** Let $X_1, \cdots, X_n$ be a random sample from P_θ, $\theta \in \Theta \subseteq \mathbb{R}^k$, and

$$\hat{\theta}_n^{MCE} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \rho(X_i, \theta).$$

Assume the following along with $E_{\theta_0}|\rho(X_1, \theta)| < \infty\ \forall \theta_0, \theta \in \Theta$:

For a *fixed* $\theta_0 \in \Theta$, $\exists$a compact set $K \subseteq \Theta$ containing θ_0 such that

(i) (Unique minimizer) $\theta_0 = \arg\min_{\theta \in K} E_{\theta_0}\rho(X_1, \theta)$, and θ_0 is the unique minimizer.

(ii) (Uniform convergence) $\sup_{\theta \in K} |\overline{\rho}_n(\theta) - E_{\theta_0}\rho(X_1, \theta)| \xrightarrow[n\to\infty]{P_{\theta_0}} 0$.

(iii) (K instead of Θ) $P_{\theta_0}(\hat{\theta}_n^{MCE} \in K) \xrightarrow[n\to\infty]{} 1$.

(iv) (Continuous D_0) A function $\theta \mapsto E_{\theta_0}\rho(X_1, \theta)$ is continuous on K.

In here,

$$\overline{\rho}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \rho(X_i, \theta).$$

Then $\hat{\theta}_n^{MCE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$.

Proof. Note that Θ need not be compact. Thus, we may use K instead of Θ. By (the proof of) theorem 1.1.24, we get

$$P_{\theta_0}\left[|\hat{\theta}_n^{MCE} - \theta_0| > \epsilon,\ \hat{\theta}_n^{MCE} \in K\right] \xrightarrow[n\to\infty]{} 0.$$

Thus, we get

$$P_{\theta_0}\left[|\hat{\theta}_n^{MCE}-\theta_0|>\epsilon\right]\le P_{\theta_0}\left[|\hat{\theta}_n^{MCE}-\theta_0|>\epsilon,\ \hat{\theta}_n^{MCE}\in K\right]+P_{\theta_0}\left[\hat{\theta}_n^{MCE}\notin K\right]\xrightarrow[n\to\infty]{}0.$$

Remark 1.1.31. Indeed, we did not see consistency of MCE yet, but we only verified for fixed $\theta_0\in\Theta$. For the consistency of MCE, we need that *for any $\theta_0\in\Theta$ $\exists K\subseteq\Theta$ containing θ_0 such that the conditions (i)-(iv) are fulfilled.* Suppose that

(a) *for all* compact $K\subseteq\Theta$ and for all $\theta_0\in\Theta$,

$$\sup_{\theta\in K}|\overline{\rho}_n(\theta)-E_{\theta_0}\rho(X_1,\theta)|\xrightarrow[n\to\infty]{P_{\theta_0}}0.$$

(b) *for any* $\theta_0\in\Theta$ there exists a compact subset K of Θ containing θ_0 such that

$$P_{\theta_0}\left(\inf_{\theta\in K^c}(\overline{\rho}_n(\theta)-\overline{\rho}_n(\theta_0))>0\right)\xrightarrow[n\to\infty]{}1.$$

(c) $\theta\mapsto E_{\theta_0}\rho(X_1,\theta)$ is continuous on K.

Then *for any* $\theta_0\in\Theta$ there exists a compact subset K of Θ containing θ_0 such that (ii)-(iv) hold. Note that, (b) implies (iii) with (i) and (c).

Also note that, MLE is a special case for MCE, $\rho(x,\theta)=-\log f(x,\theta)$.

Remark 1.1.32. In many cases, it's difficult to verify uniform convergence condition. For this, following **convexity lemma** is useful: *If K is convex,*

$$\overline{\rho}_n(\theta)\xrightarrow[n\to\infty]{P_{\theta_0}}E_{\theta_0}\rho(X_1,\theta)\ \forall\theta\in K, \qquad\text{(“pointwise convergence”)}$$

and $\overline{\rho}_n$ is a convex function on K with P_{θ_0}-a.s., then we get “uniform convergence”

$$\sup_{\theta\in K}|\overline{\rho}_n(\theta)-E_{\theta_0}\rho(X_1,\theta)|\xrightarrow[n\to\infty]{P_{\theta_0}}0.$$

See D. Pollard (1991), Econometric Theory, 7, 186-199.

Remark 1.1.33. The condition (b) in remark 1.1.31 is satisfied if the empirical contrast

$$\rho_n(\theta)=\frac{1}{n}\sum_{i=1}^n\rho(X_i,\theta)$$

is convex on a convex open parameter space $\Theta \subseteq \mathbb{R}^k$, and approaches $+\infty$ on the boundary with probability tending to 1.

1.2 The Delta Method

Basic intuition of the Delta Method is Taylor expansion.

Theorem 1.2.1. *Suppose* $\sqrt{n}(X_n - a) \xrightarrow[n\to\infty]{d} X$. *Then*

$$\sqrt{n}(g(X_n) - g(a)) = \dot{g}(a)\sqrt{n}(X_n - a) + o_P(1)$$

and hence

$$\sqrt{n}(g(X_n) - g(a)) \xrightarrow[n\to\infty]{d} \dot{g}(a)X,$$

provided g *is differentiable at* a.

Proof. By Taylor theorem, $\exists R(x,a)$ s.t.

$$g(x) = g(a) + (\dot{g}(a) + R(x,a))(x-a)$$

where $R(x,a) \to 0$ as $x \to a$. Note that if $X_n \xrightarrow[n\to\infty]{P} a$ and $R(x,a) \to 0$ as $x \to a$ then $R(X_n, a) \xrightarrow[n\to\infty]{P} 0$ ($\because \forall\epsilon > 0\ \exists\delta > 0\ s.t. |x-a| < \delta \Rightarrow |R(x,a)| < \epsilon$ implies

$$P(|R(X_n,a)| > \epsilon) \le P(|X_n - a| \ge \delta) \xrightarrow[n\to\infty]{} 0$$

and then $R(X_n, a) = o_P(1)$). Thus

$$g(X_n) = g(a) + (\dot{g}(a) + R(X_n,a))(X_n - a)$$

and hence

$$\sqrt{n}(g(X_n) - g(a)) = \dot{g}(a)\sqrt{n}(X_n - a) + \underbrace{R(X_n,a)}_{=o_P(1)}\underbrace{\sqrt{n}(X_n - a)}_{=O_P(1)} = \dot{g}(a)\sqrt{n}(X_n - a) + o_P(1).$$

In multivariate case, statement becomes $\dot{g}(a)^\top(X_n - a)$. $\square$

Remark 1.2.2. When g is a function of several variables, the differentiability means the total differentiability, which is implied by the existence of "continuous partial derivatives."

Example 1.2.3. $(X_1, Y_1), \cdots, (X_n, Y_n)$: iid from $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Let

$$\hat{\rho}_n = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

(i) As far as the distribution of $\hat{\rho}_n$ is concerned, we may assume $\mu_1 = \mu_2 = 0,\ \sigma_1^2 = \sigma_2^2 = 1$. Let

$$W_i = (X_i, Y_i, X_i^2, Y_i^2, X_iY_i)^\top \overset{i.i.d.}{\sim} (0, 0, 1, 1, \rho),$$

and $Z_n = \sqrt{n}(\overline{W}_n - (0, 0, 1, 1, \rho)^\top)$, i.e.,

$$Z_{n1} = \sqrt{n}\overline{X},\ Z_{n2} = \sqrt{n}\overline{Y},\ Z_{n3} = \sqrt{n}(\overline{X^2} - 1),\ Z_{n4} = \sqrt{n}(\overline{Y^2} - 1),\ Z_{n5} = \sqrt{n}(\overline{XY} - \rho).$$

Note that $Z_n = O_P(1)$. Then

$$\begin{aligned}
\hat{\rho}_n &= \frac{\frac{1}{\sqrt{n}}Z_{n5} + \rho - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)\left(\frac{1}{\sqrt{n}}Z_{n2}\right)}{\sqrt{1 + \frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2}\sqrt{1 + \frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2}} \\
&= \left(\frac{1}{\sqrt{n}}Z_{n5} + \rho - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)\left(\frac{1}{\sqrt{n}}Z_{n2}\right)\right) \\
&\cdot \left(1 + \frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right)^{-1/2}\left(1 + \frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right)^{-1/2} \\
&= \left(\frac{1}{\sqrt{n}}Z_{n5} + \rho - o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&\cdot \left(1 - \frac{1}{2}\left(\frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\left(\frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \left(\frac{1}{\sqrt{n}}Z_{n5} + \rho - o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\frac{1}{\sqrt{n}}Z_{n3} + o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\frac{1}{\sqrt{n}}Z_{n4} + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \left(\frac{1}{\sqrt{n}}Z_{n5} + \rho - o_P\left(\frac{1}{\sqrt{n}}\right)\right)\left(1 - \frac{1}{2}\frac{1}{\sqrt{n}}Z_{n3} - \frac{1}{2}\frac{1}{\sqrt{n}}Z_{n4} + o_P\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \rho + \frac{1}{\sqrt{n}}Z_{n5} - \frac{\rho}{2}\left(\frac{1}{\sqrt{n}}Z_{n3} + \frac{1}{\sqrt{n}}Z_{n4}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

holds, so we get

$$\sqrt{n}(\hat{\rho}_n - \rho) = Z_{n5} - \frac{\rho}{2}Z_{n3} - \frac{\rho}{2}Z_{n4} + o_P(1)$$

$$
\begin{aligned}
&= \sqrt{n}\left((\overline{XY}-\rho)-\frac{\rho}{2}(\overline{X^2}-1)-\frac{\rho}{2}(\overline{Y^2}-1)\right)+o_P(1)\\
&= \sqrt{n}\left(\overline{XY}-\frac{\rho}{2}\overline{X^2}-\frac{\rho}{2}\overline{Y^2}\right)+o_P(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(X_iY_i-\frac{\rho}{2}X_i^2-\frac{\rho}{2}Y_i^2\right)+o_P(1)\\
&\xrightarrow[n\to\infty]{d} N(0, Var\left(X_1Y_1-\frac{\rho}{2}X_1^2-\frac{\rho}{2}Y_1^2\right)).
\end{aligned}
$$

(ii) Now additionally suppose that (X_i, Y_i)'s are from bivariate normal distribution. Then $Y_1-\rho X_1$ is independent of X_1. Letting $Z_1=Y_1-\rho X_1$, we get $Var(Z_1)=1-\rho^2$, $Var(Z_1^2)=2(1-\rho^2)^2$ and hence

$$
\begin{aligned}
Var\left(X_1Y_1-\frac{\rho}{2}X_1^2-\frac{\rho}{2}Y_1^2\right) &= Var\left((1-\rho^2)X_1Z_1-\frac{\rho}{2}Z_1^2+\frac{\rho}{2}(1-\rho^2)X_1^2\right)\\
&= Var\left(\frac{\rho}{2}(1-\rho^2)X_1^2-\frac{\rho}{2}Z_1^2\right)+Var\left((1-\rho^2)X_1Z_1\right)\\
&+2\underbrace{Cov\left(\frac{\rho}{2}(1-\rho^2)X_1^2-\frac{\rho}{2}Z_1^2,(1-\rho^2)X_1Z_1\right)}_{=0}\\
&= \frac{\rho^2}{4}\left((1-\rho^2)^2Var(X_1^2)-2(1-\rho^2)Cov(X_1^2,Z_1^2)+Var(Z_1^2)\right)\\
&+(1-\rho^2)^2Var(X_1Z_1)\\
&= \frac{\rho^2}{4}\left(2(1-\rho^2)^2+2(1-\rho^2)^2\right)+(1-\rho^2)^2(1-\rho^2)\\
&= \rho^2(1-\rho^2)^2+(1-\rho^2)^2(1-\rho^2)\\
&= (1-\rho^2)^2
\end{aligned}
$$

holds. It implies that

$$\sqrt{n}(\hat{\rho}_n-\rho)\xrightarrow[n\to\infty]{d} N(0,(1-\rho^2)^2).$$

Therefore, if we define $h(\rho)$ as $h'(\rho)=(1-\rho^2)^{-1}$, i.e.,

$$h(\rho)=\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \qquad \text{(“Fisher's } z\text{-transform”)}$$

then we get

$$\sqrt{n}(h(\hat{\rho}_n)-h(\rho))\xrightarrow[n\to\infty]{d} N(0,1),$$

and with this, we can find a confidence region “with stabilized variance.”

We can also expand with higher order terms.

Theorem 1.2.4 (Higher order stochastic expansion). *Let $X_1, \cdots, X_n$ be a random sample with $EX_1 = \mu$ and finite $Var(X_1) = \Sigma$.*

(a) (1-dim case) For g with $\exists \ddot{g}$,

$$g(\overline{X}_n) = g(\mu) + \frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2 + o_P\left(\frac{1}{n}\right),$$

where $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$.

(b) (general case) For g with $\exists \ddot{g}$,

$$g(\overline{X}_n) = g(\mu) + \dot{g}(\mu)^\top(\overline{X}_n - \mu) + \frac{1}{2}(\overline{X}_n - \mu)^\top\ddot{g}(\mu)(\overline{X}_n - \mu) + o_P\left(\frac{1}{n}\right).$$

Proof. Again, use Taylor theorem. Only prove (a). Note that

$$g(x) = g(a) + \dot{g}(a)(x-a) + \frac{1}{2}\left(\ddot{g}(a) + R(x,a)\right)(x-a)^2$$

for $R(x,a) \to 0$ as $x \to a$, so letting $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$, we get

$$\begin{aligned} g(\overline{X}_n) &= g(\mu) + \dot{g}(\mu)(\overline{X}_n - \mu) + \frac{1}{2}\ddot{g}(\mu)(\overline{X}_n - \mu)^2 + \frac{1}{2}\underbrace{R(\overline{X}_n, \mu)}_{=o_P(1)}\underbrace{(\overline{X}_n - \mu)^2}_{=O_P(1/n)} \\ &= g(\mu) + \frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2 + o_P(1/n) \end{aligned}$$

which implies the conclusion.

Remark 1.2.5. For general case, following notation is also frequently used. For $\left(Z_n^i\right)_{i=1}^d = \sqrt{n}(\overline{X}_n - \mu)$,

$$g(\overline{X}_n) = g(\mu) + \frac{1}{\sqrt{n}}g_{/i}(\mu)Z_n^i + \frac{1}{2n}g_{/ij}(\mu)Z_n^iZ_n^j + o_P\left(\frac{1}{n}\right).$$

In here, we omit the "$\sum$," i.e.,

$$g_{/i}(\mu)Z_n^i := \sum_{i=1}^d g_{/i}(\mu)Z_n^i,\ g_{/ij}(\mu)Z_n^iZ_n^j = \sum_{i=1}^d\sum_{j=1}^d g_{/ij}(\mu)Z_n^iZ_n^j.$$

Example 1.2.6 (Estimation of Reliability). Let $X_1, \cdots, X_n$ be a random sample from $Exp(\lambda)$, where $\lambda > 0$ is a rate.

(i) Note that

$$\hat{\eta}_n^{MLE} = e^{-a/\overline{X}}.$$

Let $Z_n = \sqrt{n}(\lambda\overline{X} - 1)$. Then $Z_n \xrightarrow[n\to\infty]{d} N(0,1)$ and

$$\overline{X}^{-1} = \lambda\left(\frac{Z_n}{\sqrt{n}} + 1\right)^{-1}.$$

Thus, we get

$$\begin{aligned}
\hat{\eta}_n^{MLE} &= \exp\left(-a\lambda\left(\frac{Z_n}{\sqrt{n}} + 1\right)^{-1}\right)\\
&= \exp\left(-a\lambda\left(1 - \frac{Z_n}{\sqrt{n}} + \frac{Z_n^2}{n} + o_P\left(\frac{1}{n}\right)\right)\right)\\
&= e^{-a\lambda}\left(1 + \left(a\lambda\frac{Z_n}{\sqrt{n}} - a\lambda\frac{Z_n^2}{n}\right) + \frac{1}{2}\left(a\lambda\frac{Z_n}{\sqrt{n}} - a\lambda\frac{Z_n^2}{n}\right)^2 + o_P\left(\frac{1}{n}\right)\right)\\
&= e^{-a\lambda}\left(1 + a\lambda\frac{Z_n}{\sqrt{n}} + \frac{-a\lambda + (a\lambda)^2/2}{n}Z_n^2 + o_P\left(\frac{1}{n}\right)\right)\\
&= \eta + \frac{a\lambda e^{-a\lambda}}{\sqrt{n}}Z_n + \frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n}Z_n^2 + o_P\left(\frac{1}{n}\right) \qquad (1.2)
\end{aligned}$$

from $(1+x)^{-1} = 1 - x + x^2 + o(x^2)$ and $e^x = 1 + x + x^2/2 + o(x^2)$ when $x \approx 0$.

(ii) Now consider

$$\hat{\eta}^{UMVUE} = \left(1 - \frac{a}{n\overline{X}}\right)^{-1} I\left(\frac{a}{n\overline{X}} < 1\right).$$

Note that from $P\left(\frac{a}{n\overline{X}} < 1\right) \xrightarrow[n\to\infty]{} 1$, we can let $\hat{\eta}^{UMVUE} = \left(1 - \frac{a}{n\overline{X}}\right)^{-1}$ (See next remark). Then

$$\begin{aligned}
\log\hat{\eta}^{UMVUE} &= (n-1)\log\left(1 - \frac{a}{n\overline{X}}\right)\\
&= (n-1)\log\left(1 - \frac{a\lambda}{n}\left(\frac{Z_n}{\sqrt{n}} + 1\right)^{-1}\right)\\
&= (n-1)\log\left(1 - \frac{a\lambda}{n}\left(1 - \frac{Z_n}{\sqrt{n}} + \frac{Z_n^2}{n} + o_P\left(\frac{1}{n}\right)\right)\right)\\
&= (n-1)\log\left(1 - \frac{a\lambda}{n} + \frac{a\lambda}{n\sqrt{n}}Z_n - \frac{a\lambda}{n^2}Z_n^2 + o_P\left(\frac{1}{n^2}\right)\right)\\
&= (n-1)\left\{\left(-\frac{a\lambda}{n} + \frac{a\lambda}{n\sqrt{n}}Z_n - \frac{a\lambda}{n^2}Z_n^2\right) - \frac{1}{2}\left(-\frac{a\lambda}{n} + \frac{a\lambda}{n\sqrt{n}}Z_n - \frac{a\lambda}{n^2}Z_n^2\right)^2\right\} + o_P\left(\frac{1}{n}\right)\\
&= -a\lambda + \frac{a\lambda}{\sqrt{n}}Z_n - \frac{a\lambda}{n}Z_n^2 - \frac{(a\lambda)^2}{2n} + \frac{a\lambda}{n} + o_P\left(\frac{1}{n}\right)\\
&= -a\lambda + \frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n} + o_P\left(\frac{1}{n}\right)
\end{aligned}$$

implies

$$\begin{aligned}
\hat{\eta}^{UMVUE} &= \exp\left(-a\lambda + \frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n} + o_P\left(\frac{1}{n}\right)\right) \\
&= e^{-a\lambda}\left(1 + \left(\frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n}\right) + \frac{1}{2}\left(\frac{a\lambda}{\sqrt{n}}Z_n + \frac{-a\lambda Z_n^2 + a\lambda - (a\lambda)^2/2}{n}\right)^2\right) \\
&+ o_P\left(\frac{1}{n}\right) \\
&= e^{-a\lambda}\left(1 + \frac{a\lambda}{\sqrt{n}}Z_n + \left(-a\lambda Z_n^2 + a\lambda - \frac{(a\lambda)^2}{2} + \frac{1}{2}(a\lambda)^2 Z_n^2\right)\frac{1}{n}\right) + o_P\left(\frac{1}{n}\right) \\
&= \eta + \frac{a\lambda e^{-a\lambda}}{\sqrt{n}}Z_n + \frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n}(Z_n^2 - 1) + o_P\left(\frac{1}{n}\right) \qquad (1.3)
\end{aligned}$$

from $\log(1+x) = x - x^2/2 + o(x^2),\ x \approx 0$. Comparing (1.3) to (1.2), we can say that UMVUE is "closer" than MLE to η, since MLE's leading term has a bias

$$\frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n},$$

while UMVUE's leading term has no bias. Like this case, if one suggests a new estimator, then in many cases, one compares 2nd order term to judge its asymptotic behavior.

Remark 1.2.7. If there is an event that occurring probability converges to 1, then in an asymptotic sense, we may ignore such event, in the sense that:

(i) If $P(\mathcal{E}_n) \xrightarrow[n\to\infty]{} 1$ and $P(X_n \le x,\ \mathcal{E}_n) \xrightarrow[n\to\infty]{} F(x)$, then $X_n \xrightarrow[n\to\infty]{d} F$.

(ii) If $P(\mathcal{E}_n) \xrightarrow[n\to\infty]{} 1$ and $X_n = X + O_P(n^{-\alpha})$ on $\mathcal{E}_n$, then $X_n = X + O_P(n^{-\alpha})$ in general. Convergence rate of $P(\mathcal{E}_n)$ does not matter!

($\because$ $P(n^\alpha|X_n - X| \ge C) \le P(n^\alpha|X_n - X| \ge C,\ \mathcal{E}_n) + \underbrace{P(\mathcal{E}_n^c)}_{\xrightarrow[n\to\infty]{} 0}$, take $\limsup_n$ and $\lim_C$ on both sides.)

Example 1.2.8. Consider a sample correlation coefficient again. Assume $EX_1^4 < \infty$ and $EY_1^4 < \infty$. Then

$$\begin{aligned}
\hat{\rho}_n &= \left(\frac{1}{\sqrt{n}}Z_{n5} + \rho - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)\left(\frac{1}{\sqrt{n}}Z_{n2}\right)\right) \\
&\cdot \left(1 + \frac{1}{\sqrt{n}}Z_{n3} - \left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right)^{-1/2}\left(1 + \frac{1}{\sqrt{n}}Z_{n4} - \left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right)^{-1/2} \\
&= \left(\rho + \frac{1}{\sqrt{n}}Z_{n5} - \frac{1}{n}Z_{n1}Z_{n2} + o_P\left(\frac{1}{n}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&\cdot\left(1-\frac{1}{2}\left(\frac{1}{\sqrt{n}}Z_{n3}-\left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right)+\frac{3}{8}\left(\frac{1}{\sqrt{n}}Z_{n3}-\left(\frac{1}{\sqrt{n}}Z_{n1}\right)^2\right)^2+o_P\left(\frac{1}{n}\right)\right)\\
&\cdot\left(1-\frac{1}{2}\left(\frac{1}{\sqrt{n}}Z_{n4}-\left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right)+\frac{3}{8}\left(\frac{1}{\sqrt{n}}Z_{n4}-\left(\frac{1}{\sqrt{n}}Z_{n2}\right)^2\right)^2+o_P\left(\frac{1}{n}\right)\right)\\
&=\left(\rho+\frac{1}{\sqrt{n}}Z_{n5}-\frac{1}{n}Z_{n1}Z_{n2}+o_P\left(\frac{1}{n}\right)\right)\\
&\cdot\left(1-\frac{1}{2\sqrt{n}}Z_{n3}+\frac{1}{n}\left(\frac{1}{2}Z_{n1}^2+\frac{3}{8}Z_{n3}^2\right)+o_P\left(\frac{1}{n}\right)\right)\\
&\cdot\left(1-\frac{1}{2\sqrt{n}}Z_{n4}+\frac{1}{n}\left(\frac{1}{2}Z_{n2}^2+\frac{3}{8}Z_{n4}^2\right)+o_P\left(\frac{1}{n}\right)\right)\\
&=\rho+\frac{1}{\sqrt{n}}\left(Z_{n5}-\frac{\rho}{2}Z_{n3}-\frac{\rho}{2}Z_{n4}\right)\\
&+\frac{1}{n}\left(-Z_{n1}Z_{n2}-\frac{1}{2}Z_{n3}Z_{n5}-\frac{1}{2}Z_{n4}Z_{n5}+\rho\left(\frac{1}{4}Z_{n3}Z_{n4}+\frac{1}{2}Z_{n1}^2+\frac{1}{2}Z_{n2}^2+\frac{3}{8}Z_{n3}^2+\frac{3}{8}Z_{n4}^2\right)\right)+o_P\left(\frac{1}{n}\right)
\end{aligned}
$$

holds. The leading term has bias

$$\frac{1}{n}\left\{\frac{\rho}{4}EX_1^2Y_1^2+\frac{3}{8}\rho(EX_1^4+EY_1^4)-\frac{1}{2}(EX_1^3Y_1+EX_1Y_1^3)\right\}$$

from

$$
\begin{aligned}
E(Z_{3n}Z_{4n})&=EX_1^2X_2^2-1\\
E(Z_{3n}^2+Z_{4n}^2)&=EX_1^4+EY_1^4-2\\
E(Z_{1n}^2+Z_{2n}^2)&=2\\
E(Z_{3n}Z_{5n}+Z_{4n}Z_{5n})&=EX_1^3Y_1+EX_1Y_1^3-2\rho\\
E(Z_{1n}Z_{2n})&=\rho.
\end{aligned}
$$

For bivariate normal case, it becomes

$$-\frac{1}{2n}\rho(1-\rho^2).$$

Remark 1.2.9. Note that in using stochastic expansion, we can get the mean, variance, skewness, ... of the leading term and might expect that they become the approximation of the moments of $g(\overline{X}_n)$, but this is not true! For example, if $X_n \sim Ber(1/n)$, then

$$P(nX_n>\epsilon)=P(X_n=1)=\frac{1}{n}\xrightarrow[n\to\infty]{}0$$

from $X_n = 0$ or 1, so $nX_n = o_P(1)$, but we get $E(nX_n) = 1\ \forall n$. Thus, we need to check the behavior of the remainder.

From now on, we compare "moments of leading terms" and "approximation of moments."

Example 1.2.10. Recall mgf

$$mgf_X(t) = Ee^{tX} \ \ |t| < \epsilon$$

and cgf

$$cgf_X(t) = \log mgf_X(t) = \log Ee^{tX}. \ |t| < \epsilon$$

For $m_r = EX^r$, mgf has a Taylor expansion

$$mgf_X(t) = 1 + m_1 t + \frac{m_2}{2!}t^2 + \frac{m_3}{3!}t^3 + \cdots ,$$

and if X and Y are independent,

$$mgf_{X+Y}(t) = mgf_X(t) \cdot mgf_Y(t)$$

and

$$cgf_{aX+b}(t) = cgf_X(at) + bt$$

holds for constants a and b. From this we get

$$c_r(aX + b) = a^r c_r(X),$$

where c_r denotes rth cumulant. Also recall that, for $A \approx 0$,

$$\log(1 + A) = A - \frac{1}{2}A^2 + \frac{1}{3}A^3 - \cdots ,$$

and with this, we can obtain

$$cgf_X(t) = \log\left(1 + \underbrace{(mgf_X(t) - 1)}_{=A}\right) = c_1 t + \frac{c_2}{2!}t^2 + \frac{c_3}{3!}t^3 + \cdots$$

where

$$c_1 = m_1,\ c_2 = m_2 - m_1^2,\ c_3 = m_3 - 3m_1m_2 + 2m_1^3,\ c_4 = m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - m_1^4, \cdots .$$

If observations are normalized, i.e., $m_1 = 0$ and $m_2 = 1$, then

$$c_1 = 0,\ c_2 = 1,\ c_3 = m_3,\ c_4 = m_4 - 3.$$

Example 1.2.11. Let

$$Z_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \underbrace{\frac{X_i - \mu}{\sigma}}_{=:\tilde{Z}_i \stackrel{d}{\equiv} Z_1}.$$

Then from

$$cgf_{Z_n}(t) = cgf_{n^{-1/2}\sum \tilde{Z}_i}(t) = n \cdot cgf_{Z_1}\left(\frac{t}{\sqrt{n}}\right),$$

we obtain

$$c_r(Z_n) = n \cdot \left(\frac{1}{\sqrt{n}}\right)^r c_r(Z_1) = n^{-\frac{r}{2}+1} c_r(Z_1).$$

From this, we obtain

$$EZ_n^3 = c_3(Z_n) = \frac{1}{\sqrt{n}} c_3(Z_1) \tag{1.4}$$

$$EZ_n^4 = c_4(Z_n) + 3 = \frac{1}{n} c_4(Z_1) + 3. \tag{1.5}$$

Example 1.2.12. Now see the multivariate case. Let $X = (X_1, \cdots, X_d)^\top$ and $t = (t_1, \cdots, t_d)^\top$. Then

$$mgf_X(t) = Ee^{t^\top X} = Ee^{t_1 X_1 + \cdots + t_d X_d}$$

and

$$m_1 = \left[\frac{\partial}{\partial t_i} mgf_X(t)\Big|_{t=0}\right]_i$$

$$m_2 = \left[\frac{\partial^2}{\partial t_i \partial t_j} mgf_X(t)\Big|_{t=0}\right]_{i,j}$$

$$m_3 = \left[\frac{\partial^3}{\partial t_i \partial t_j \partial t_k} mgf_X(t)\Big|_{t=0}\right]_{i,j,k}$$

$$\vdots$$

and we get

$$mgf_X(t) = 1 + \sum_i m_1(i) t_i + \frac{1}{2!}\sum_{i,j} m_2(i,j) t_i t_j + \frac{1}{3!}\sum_{i,j,k} m_3(i,j,k) t_i t_j t_k + \cdots.$$

If data is centered, i.e., $EX = 0$, then $m_1 = 0$ and so

$$cgf_X(t) = \frac{1}{2!}\sum_{i,j} m_2(i,j)t_it_j + \frac{1}{3!}\sum_{i,j,k} m_3(i,j,k)t_it_jt_k + \frac{1}{4!}\sum_{i,j,k,l}\left(m_4(i,j,k,l) - 3m_2(i,j)m_2(k,l)\right)t_it_jt_kt_l + \cdots .$$

Now let $Z_n = (Z_n^1, \cdots, Z_n^d)^\top = \sqrt{n}(\overline{X}_n - \mu)$. Then

$$cgf_{Z_n}(t) = n \cdot cgf_{Z_1}(t/\sqrt{n})$$

implies

$$EZ_n^iZ_n^j =: \sigma^{i,j},\ (\sigma^{i,j})_{i,j} = Var(X_1)$$

$$EZ_n^iZ_n^jZ_n^k = \frac{1}{\sqrt{n}}c_3(i,j,k)$$

$$EZ_n^iZ_n^jZ_n^kZ_n^l = \frac{1}{n}c_4(i,j,k,l) + 3\sigma^{ij}\sigma^{kl}$$

where

$$c_3(i,j,k) = c_3(Z_1)(i,j,k) = EZ_1^iZ_1^jZ_1^k$$

$$c_4(i,j,k,l) = c_4(Z_1)(i,j,k,l) = EZ_1^iZ_1^jZ_1^kZ_1^l - 3(EZ_1^iZ_1^j)(EZ_1^kZ_1^l).$$

Proposition 1.2.13 (Moments of the leading terms). *Let $X_1, \cdots, X_n$ be i.i.d. with $E|X_1|^4 < \infty$*[1]*.*

(a) (Univariate case) For g with $\exists \ddot{g}$ and $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$,

$$g(\overline{X}_n) = \underbrace{g(\mu) + \frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2}_{=:W_n} + o_P(n^{-1}),$$

and for the leading term W_n,

$$\begin{aligned}
E(W_n) &= g(\mu) + \frac{\sigma^2}{2n}\ddot{g}(\mu)\\
Var(W_n) &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 + \frac{1}{n^2}\left(\sigma^3c_3(Z_1)\dot{g}(\mu)\ddot{g}(\mu) + \frac{1}{2}\sigma^4\left(\ddot{g}(\mu)\right)^2\right) + O(n^{-3})\\
E(W_n - EW_n)^3 &= \frac{1}{n^2}\left(\sigma^3c_3(Z_1)\left(\dot{g}(\mu)\right)^3 + 3\sigma^4\left(\dot{g}(\mu)\right)^2\left(\ddot{g}(\mu)\right)\right) + O(n^{-3}).
\end{aligned}$$

[1]In fact, stronger condition is needed: mgf of X_1 exists

(b) (Multivariate case) For g with $\exists\ddot{g}$ and $(Z_n^i) = \sqrt{n}(\overline{X}_n - \mu)$,

$$g(\overline{X}_n) = \underbrace{g(\mu) + \frac{1}{\sqrt{n}}g_{/i}(\mu)Z_n^i + \frac{1}{2n}g_{/ij}(\mu)Z_n^i Z_n^j}_{=:W_n} + o_P(n^{-1}),$$

and for the leading term W_n,

$$\begin{aligned}E(W_n) &= g(\mu) + \frac{1}{2n}g_{/ij}(\mu)\sigma^{ij}\\ Var(W_n) &= \frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\sigma^{ij} + O(n^{-2})\end{aligned}$$

where $(\sigma^{ij}) = Var(X_1)$.

Proof. (a) Nothing but tedious calculation. First,

$$E(W_n) = g(\mu) + \frac{\sigma^2}{2n}\ddot{g}(\mu)$$

is easily obtained. Next, note that

$$\begin{aligned}Var(W_n) &= Var\left(\frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)Z_n^2\right)\\ &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 Var(Z_n) + \frac{\sigma^3}{n\sqrt{n}}\dot{g}(\mu)\ddot{g}(\mu)Cov(Z_n, Z_n^2) + \frac{\sigma^4}{4n^2}\left(\ddot{g}(\mu)\right)^2 Var(Z_n^2).\end{aligned}$$

First, $Var(Z_n) = 1$. Also, $Var(Z_n^2) = E(Z_n^4) - \left[E(Z_n^2)\right]^2 = 2 + n^{-1}c_4(Z_1)$ from (1.5). Finally, $Cov(Z_n, Z_n^2) = EZ_n^3 - EZ_n \cdot EZ_n^2 = n^{-1/2}c_3(Z_1)$ from (1.4). Now we get

$$\begin{aligned}Var(W_n) &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 + \frac{\sigma^3}{n^2}\dot{g}(\mu)\ddot{g}(\mu)c_3(Z_1) + \frac{\sigma^4}{2n^2}\left(\ddot{g}(\mu)\right)^2 + \frac{\sigma^4}{4n^3}\left(\ddot{g}(\mu)\right)^2 c_4(Z_1)\\ &= \frac{\sigma^2}{n}\left(\dot{g}(\mu)\right)^2 + \frac{1}{n^2}\left(\sigma^3 c_3(Z_1)\dot{g}(\mu)\ddot{g}(\mu) + \frac{1}{2}\sigma^4\left(\ddot{g}(\mu)\right)^2\right) + O(n^{-3}).\end{aligned}$$

For $E(W_n - EW_n)^3$, note that $EZ_n^5 = O(n^{-1/2})$ and $EZ_n^6 = O(1)$. (Check!)[2] Then

$$\begin{aligned}E(W_n - EW_n)^3 &= E\left[\frac{\sigma}{\sqrt{n}}\dot{g}(\mu)Z_n + \frac{\sigma^2}{2n}\ddot{g}(\mu)(Z_n^2 - 1)\right]^3\\ &= \frac{\sigma^3}{n\sqrt{n}}\dot{g}(\mu)^3 \underbrace{EZ_n^3}_{=n^{-1/2}c_3(Z_1)} + \frac{3\sigma^4}{2n^2}\dot{g}(\mu)^2\ddot{g}(\mu)\underbrace{E[Z_n^2(Z_n^2 - 1)]}_{=n^{-1}c_4(Z_1)+2}\end{aligned}$$

[2]For EZ_n^5, check directly, and for EZ_n^6, you can check directly, or use $Z_n^6 = O_P(1)$ from $Z_n = O_P(1)$.

$$
\begin{aligned}
&+\underbrace{\frac{\sigma^5}{4n^2\sqrt{n}}\dot{g}(\mu)\ddot{g}(\mu)^2E[Z_n(Z_n^2-1)^2]+\frac{\sigma^6}{8n^3}\ddot{g}(\mu)^3E[(Z_n^2-1)^3]}_{=O(n^{-3})}\\
&=\frac{1}{n^2}\left(\sigma^3c_3(Z_1)\left(\dot{g}(\mu)\right)^3+3\sigma^4\left(\dot{g}(\mu)\right)^2\left(\ddot{g}(\mu)\right)\right)+O(n^{-3})
\end{aligned}
$$

by (1.4) and (1.5).

(b) Note that

$$
W_n=g(\mu)+\frac{1}{\sqrt{n}}g_{/i}(\mu)Z_n^i+\frac{1}{2n}g_{/ij}(\mu)Z_n^iZ_n^j=g(\mu)+\frac{1}{\sqrt{n}}\sum_{i=1}^d g_{/i}(\mu)Z_n^i+\frac{1}{2n}\sum_{i,j=1}^d g_{/ij}(\mu)Z_n^iZ_n^j
$$

so

$$
\begin{aligned}
Var(W_n)&=\frac{1}{n}\sum_{i=1}^d\sum_{j=1}^d g_{/i}(\mu)g_{/j}(\mu)Cov(Z_n^i,Z_n^j)+\frac{1}{n\sqrt{n}}\sum_{i=1}^d\sum_{k,l=1}^d g_{/i}(\mu)g_{/kl}(\mu)Cov(Z_n^i,Z_n^kZ_n^l)\\
&+\frac{1}{4n^2}\sum_{i,j=1}^d\sum_{k,l=1}^d g_{/ij}(\mu)g_{/kl}(\mu)Cov(Z_n^iZ_n^j,Z_n^kZ_n^l)\\
&=\frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\underbrace{Cov(Z_n^i,Z_n^j)}_{=\sigma^{ij}}+\frac{1}{n\sqrt{n}}g_{/i}(\mu)g_{/kl}(\mu)\underbrace{Cov(Z_n^i,Z_n^kZ_n^l)}_{=EZ_n^iZ_n^kZ_n^l}\\
&+\frac{1}{4n^2}g_{/ij}(\mu)g_{/kl}(\mu)\underbrace{Cov(Z_n^iZ_n^j,Z_n^kZ_n^l)}_{EZ_n^iZ_n^jZ_n^kZ_n^l-(EZ_n^iZ_n^j)(EZ_n^kZ_n^l)}\\
&=\frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\sigma^{ij}+\frac{1}{n^2}g_{/i}(\mu)g_{/kl}(\mu)c_3(i,j,k)+\frac{1}{4n^2}g_{/ij}(\mu)g_{/kl}(\mu)\left(\frac{1}{n}c_4(i,j,k,l)+2\sigma^{ij}\sigma^{kl}\right)\\
&=\frac{1}{n}g_{/i}(\mu)g_{/j}(\mu)\sigma^{ij}+O(n^{-2}).
\end{aligned}
$$

□

Proposition 1.2.14 (Approximation of moments). *Let $X_1,\cdots,X_n$ be i.i.d. with $E|X_1|^4<\infty$.*

(a) (Univariate case) For g with bounded $g^{(r)}$ $(r=0,1,\cdots,4)$,

$$
\begin{aligned}
E(g(\overline{X}_n))&=g(\mu)+\frac{\sigma^2}{2n}g^{(2)}(\mu)+O(n^{-2})\\
Var(g(\overline{X}_n))&=\frac{\sigma^2}{n}g^{(1)}(\mu)^2+O(n^{-2})
\end{aligned}
$$

where $\mu=EX_1$, $\sigma^2=Var(X_1)$.

(b) (Multivariate case) For g with bounded and contiuous $g_{/I}$ ($|I| = 0, 1, \cdots, 4$),

$$E(g(\overline{X}_n)) = g(\mu) + \frac{1}{2n} g_{/ij}(\mu)\sigma^{ij} + O(n^{-2})$$
$$Var(g(\overline{X}_n)) = \frac{1}{n} g_{/i}(\mu) g_{/j}(\mu)\sigma^{ij} + O(n^{-2})$$

where $\mu = EX_1$, $(\sigma^{ij}) = Var(X_1)$.

Proof. (a) Note that,

$$g(\overline{X}_n) = g(\mu) + \frac{\sigma}{\sqrt{n}} g^{(1)}(\mu) Z_n + \frac{\sigma^2}{2n} g^{(2)}(\mu) Z_n^2 + \frac{1}{3!}\frac{\sigma^3}{n\sqrt{n}} g^{(3)}(\mu) Z_n^3 + R_n,$$

where

$$R_n = \frac{1}{4!}\frac{\sigma^4}{n^2} g^{(4)}(\xi_n) Z_n^4,\ \xi_n:\ \text{a number between } \mu \text{ and } \overline{X}_n.$$

Note that

$$E|R_n| \le \frac{1}{4!}\frac{\sigma^4}{n^2}\sup_x |g^{(4)}(x)| \cdot EZ_n^4 = O(n^{-2})$$

from "boundedness of $g^{(4)}$" and $EZ_n^4 = 3 + n^{-1}c_4(Z_1)$. Also note that

$$\frac{\sigma^3}{n\sqrt{n}} EZ_n^3 = \frac{\sigma^3}{n\sqrt{n}}\frac{1}{\sqrt{n}} c_3(Z_1) = O(n^{-2}).$$

From these, we obtain

$$Eg(\overline{X}_n) = g(\mu) + \frac{\sigma^2}{2n} g^{(2)}(\mu).$$

Next, for the variance, we get

$$\begin{aligned}
Var g(\overline{X}_n) &= \frac{\sigma^2}{n} g^{(1)}(\mu)^2 Var(Z_n) + \frac{\sigma^2}{4n^2} g^{(2)}(\mu)^2 Var(Z_n^2) \\
&\quad + O(n^{-3}) Var(Z_n^3) + Var(R_n) \\
&\quad + O(n^{-3/2}) Cov(Z_n, Z_n^2) + O(n^{-2}) Cov(Z_n, Z_n^3) + O(n^{-5/2}) Cov(Z_n^2, Z_n^3) \\
&\quad + O(n^{-1/2}) Cov(Z_n, R_n) + O(n^{-1}) Cov(Z_n^2, R_n) + O(n^{-3/2}) Cov(Z_n^3, R_n).
\end{aligned}$$

Note that,

$$Var(Z_n) = 1,\ Var(Z_n^2) = 2 + \frac{1}{n} c_4(Z_1) = O(1),\ Var(Z_n^3) = EZ_n^6 - (EZ_n^3)^2 \le O(1),$$

$$Var(R_n) \le ER_n^2 = O(n^{-4}) \cdot EZ_n^8 \le O(n^{-4}),$$

$$|Cov(Z_n, R_n)| \le \sqrt{VarZ_n}\sqrt{VarR_n} \le O(n^{-2}),$$

$$|Cov(Z_n^2, R_n)| \le \sqrt{VarZ_n^2}\sqrt{VarR_n} \le O(n^{-2}),$$

$$|Cov(Z_n^3, R_n)| \le \sqrt{VarZ_n^3}\sqrt{VarR_n} \le O(n^{-2}),$$

$$Cov(Z_n, Z_n^2) = EZ_n^3 - (EZ_n)(EZ_n^2) = O(n^{-1/2}),$$

$$|Cov(Z_n, Z_n^3)| \le \sqrt{VarZ_n}\sqrt{VarZ_n^3} \le O(1),$$

$$|Cov(Z_n^2, Z_n^3)| \le \sqrt{VarZ_n^2}\sqrt{VarZ_n^3} \le O(1).$$

From these, we get

$$\begin{aligned}
Varg(\overline{X}_n) = \frac{\sigma^2}{n}g^{(1)}(\mu)^2 \underbrace{Var(Z_n)}_{=1} + \frac{\sigma^2}{4n^2}g^{(2)}(\mu)^2 \underbrace{Var(Z_n^2)}_{=O(1)} \\
+ O(n^{-3})\underbrace{Var(Z_n^3)}_{=O(1)} + \underbrace{Var(R_n)}_{\le O(n^{-4})} \\
+ O(n^{-3/2})\underbrace{Cov(Z_n, Z_n^2)}_{=O(n^{-1/2})} + O(n^{-2})\underbrace{Cov(Z_n, Z_n^3)}_{\le O(1)} + O(n^{-5/2})\underbrace{Cov(Z_n^2, Z_n^3)}_{\le O(1)} \\
+ O(n^{-1/2})\underbrace{Cov(Z_n, R_n)}_{\le O(n^{-2})} + O(n^{-1})\underbrace{Cov(Z_n^2, R_n)}_{\le O(n^{-2})} + O(n^{-3/2})\underbrace{Cov(Z_n^3, R_n)}_{\le O(n^{-2})} \\
= \frac{\sigma^2}{n}g^{(1)}(\mu)^2 + O(n^{-2}).
\end{aligned}$$

(b) Recall the Taylor theorem for multivariate function: for $(k+1)$-time continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\mathbf{x}) = \sum_{|\alpha| \le k} \frac{D^\alpha f(\mathbf{a})}{\alpha!}(\mathbf{x} - \mathbf{a})^\alpha + \sum_{|\beta| = k+1} R_\beta(\mathbf{x})(\mathbf{x} - \mathbf{a})^\beta,$$

where for $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n$ we define $|\alpha| = \alpha_1 + \cdots + \alpha_n$, $\alpha! = \alpha_1! \cdots \alpha_n!$, and $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$,

$$R_\beta(\mathbf{x}) = \frac{|\beta|}{\beta!}\int_0^1 (1-t)^{|\beta|-1} D^\beta f(\mathbf{a} + t(\mathbf{x} - \mathbf{a}))dt.$$

In here,

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Using this, we obtain

$$g(\overline{X}_n) = g(\mu) + \frac{1}{\sqrt{n}} g_{/i}(\mu) Z_n^i + \frac{1}{2!}\frac{1}{n} g_{/ij} Z_n^i Z_n^j + \frac{1}{3!}\frac{1}{n\sqrt{n}} g_{/ijk} Z_n^i Z_n^j Z_n^k + R_n,$$

where

$$|R_n| \le \frac{1}{6n^2} \left| \int_0^1 (1-u)^3 g_{/ijkl}\left(\mu + \frac{1}{\sqrt{n}} u Z_n\right) Z_n^i Z_n^j Z_n^k Z_n^l du \right|.$$

By boundedness of $g_{/I}$ and

$$E|Z_n^i Z_n^j Z_n^k Z_n^l| \le \frac{1}{4} E((Z_n^i)^4 + (Z_n^j)^4 + (Z_n^k)^4 + (Z_n^l)^4) = O(1), \qquad \text{(“AM-GM”)}$$

we obtain the desired result with similar procedure as univariate case. □

Remark 1.2.15 (Open Questions)**.**

(i) In (a), we may not use boundedness of $g^{(r)}$ ($r = 0, 1, 2, 3$), but only use that of $g^{(4)}$. Why boundedness condition is given?

(ii) Expand $g(\overline{X}_n)$ by quadratic terms

$$g(\overline{X}_n) = g(\mu) + \frac{\sigma}{\sqrt{n}} g^{(1)}(\mu) Z_n + \frac{\sigma^2}{2n} g^{(2)}(\mu) Z_n^2 + R_n,$$

$$R_n = \frac{1}{3!}\frac{\sigma^3}{n\sqrt{n}} g^{(3)}(\xi_n) Z_n^3, \ \xi_n: \text{ a number between } \mu \text{ and } \overline{X}_n.$$

Then

$$E|R_n| \le O(n^{-3/2}) E|Z_n^3| = O(n^{-2})$$

from boundedness of $g^{(3)}$. Then with similar procedure, we can get the same result. Why we should consider 3rd order term?

(Sol. We cannot guarantee $E|Z_n^3| = O(n^{-1/2})$ from $EZ_n^3 = O(n^{-1/2})$!)

(iii) I think there is a typo in the lecture note; in the note, it is written that

$$R_n = \frac{1}{6n^2} \int_0^1 (1-u)^3 g_{/ijkl}\left(\frac{1}{\sqrt{n}} u Z_n\right) Z_n^i Z_n^j Z_n^k Z_n^l du.$$

Example 1.2.16 (Estimation of reliability)**.** Let $X_1, \cdots, X_n$ be a random sample from $Exp(\lambda)$, $\lambda > 0$. Define

$$\eta = P_\lambda(X_1 > a) = e^{-a\lambda}.$$

Then

$$\hat{\eta}_n^{MLE} = \exp(-a/\overline{X}).$$

Here, $g(x) = \exp(-a/x)$ is infinitely differentiable with bounded derivatives. Thus

$$E(\hat{\eta}_n^{MLE}) = \eta + \frac{(-a\lambda + (a\lambda)^2/2)e^{-a\lambda}}{n} + O(n^{-2})$$

$$Var(\hat{\eta}_n^{MLE}) = \frac{(a\lambda)^2 e^{-2a\lambda}}{n} + O(n^{-2}),$$

the leading terms of which agree with the mean and variance of the leading term in its stochastic expansion. On the other hand, for

$$\hat{\eta}_n^{UMVUE} = \left(1 - \frac{a}{n\overline{x}}\right)^{n-1} I\left(\frac{a}{n\overline{X}} < 1\right),$$

we cannot apply the result for the approximation of moments. But, it can be shown that its variance is also approximated by the same approximation formula.

Remark 1.2.17. The boundedness of the derivatives for the approximation of moments is rather stronger than needed. Whenever the approximation can be proved, the formulae agree with the moments of the leading term of its stochastic expansion. So only the validity of the order of the remainder needs to be proved. For example, in the bivariate normal case, the mean and variance of the sample correlation coefficient can be approximated as follows;

$$E(\hat{\rho}_n) = \frac{-\rho(1-\rho^2)}{2n} + O(n^{-2})$$

$$Var(\hat{\rho}_n) = \frac{(1-\rho^2)^2}{n} + O(n^{-2}).$$

1.3 Asymptotic Theory

1.3.1 MLE in Exponential Family

Proposition 1.3.1. *Let $X_1, \cdots, X_n$ be a random sample from a population with pdf*

$$p_\eta(x) = h(x)\exp(\eta^\top T(x) - A(\eta))I_{\mathcal{X}}(x),\ \ \eta \in \mathcal{E},$$

where $\mathcal{E}$ is a natural parameter space in $\mathbb{R}^k$. Further, assume that

(i) $\mathcal{E}$ is open.

(ii) The family is of rank k.

Then

(a) $\hat{\eta}_n^{MLE} = \eta + (\ddot{A}(\eta))^{-1}(\overline{T}_n - \dot{A}(\eta)) + o_{P_\eta}(n^{-1/2}) = \eta + (-\ddot{l}(\eta))^{-1}\dot{l}(\eta) + o_{P_\eta}(n^{-1/2})$.

(b) $\sqrt{n}(\hat{\eta}_n^{MLE} - \eta) \xrightarrow[n\to\infty]{d} N\left(0, (\ddot{A}(\eta))^{-1}\right) = N(0, I_1^{-1}(\eta))$.

Proof. Recall that, under the same assumptions,

$$(\overline{T}_n \in C^0) \subseteq (\dot{A}(\hat{\eta}_n^{MLE}) = \overline{T}_n) \subseteq (\hat{\eta}_n^{MLE} = \left(\dot{A}\right)^{-1}(\overline{T}_n)),$$

with the probabilities of these events tending to 1 (See theorem 1.1.20). Also note that, by full rank condition, $\dot{A}$ is one-to-one and differentiable on $\mathcal{E}$ with $\ddot{A} > 0$.

Now let $t = \dot{A}(\eta)$. Then by **Inverse Function Theorem** (see next Remark), $\dot{A}^{-1}$ is also differentiable and

$$D(\dot{A}^{-1})(t) = \frac{\partial}{\partial t}\dot{A}^{-1}(t) = \left(\ddot{A}(\eta)\right)^{-1}.$$

CLT implies

$$\sqrt{n}(\overline{T}_n - E_\eta T(X_1)) \xrightarrow[n\to\infty]{d} N\left(0, Var_\eta T(X_1)\right),$$

and recall that $E_\eta T(X_1) = \dot{A}(\eta)$, $Var_\eta T(X_1) = \ddot{A}(\eta)$. Therefore, Δ-method implies

$$\begin{aligned}
\hat{\eta}_n^{MLE} &= \dot{A}^{-1}(\overline{T}_n) \\
&= \dot{A}^{-1}(t) + \left(\ddot{A}(\eta)\right)^{-1}(\overline{T}_n - t) + o(|\overline{T}_n - t|) \\
&= \eta + \left(\ddot{A}(\eta)\right)^{-1}(\overline{T}_n - t) + o_{P_\eta}(n^{-1/2}),
\end{aligned}$$

and hence

$$\begin{aligned}
\sqrt{n}(\hat{\eta}_n^{MLE} - \eta) &= \left(\ddot{A}(\eta)\right)^{-1}\sqrt{n}(\overline{T}_n - t) + o_{P_\eta}(1) \\
&\xrightarrow[n\to\infty]{d} N\left(0, Var_\eta T(X_1)\right).
\end{aligned}$$

Rest part is obtained from $\overline{T}_n - \dot{A}(\eta) = \dot{l}_n(\eta)/n$ and $-\ddot{l}_n(\eta) = \ddot{A}(\eta)$. □

Remark 1.3.2 (Inverse Function Theorem)**.** Let $F : \mathbb{U} \to \mathbb{R}^d$, where $U \subseteq \mathbb{R}^d$ is an open set. If

(i) F is one-to-one

(ii) F is Fréchet differentiable near $x_0 \in U$

(iii) $DF_{x_0} := \left[\frac{\partial F}{\partial x_j}\bigg|_{x=x_0} \right]_{i,j}$ is invertible.

Then $F^{-1} : F(U) \to U$ is also Fréchet differentiable at $y_0 = F(x_0)$, and it satisfies $D(F^{-1})(y_0) = (DF_{x_0})^{-1}$.

Example 1.3.3 (Multinomial case)**.** Let $X_1, \cdots, X_n$ be a random sample from $\text{Multi}(1, p(\theta))$, where $p(\theta) = (p_1(\theta), \cdots, p_k(\theta))^\top$, and $\theta \in \Theta \subseteq \mathbb{R}$. For example, consider Hardy-Weinberg proportions $p(\theta) = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)^\top$, $0 < \theta < 1$. Assume that

(i) Θ is open and $0 < p_i(\theta) < 1$, $\sum_{i=1}^k p_i(\theta) = 1$.

(ii) $p(\theta) = (p_1(\theta), \cdots, p_k(\theta))^\top$ is twice (totally) differentiable.

Then we can derive the asymptotic distribution of estimator of θ. Let $\theta = h(p(\theta))$ for any $\theta \in \Theta$ for some differentiable function h. Let

$$\hat{p}(\theta) = \frac{1}{n}\sum_{i=1}^n X_i = \left(\frac{N_1}{n}, \cdots, \frac{N_k}{n}\right)^\top,$$

where $N_j = \sum_{i=1}^n I(X_{ij} = 1)$. Then we get

$$E_\theta \hat{p}(\theta) = p(\theta)$$

and hence

$$h(\overline{X}_n) = h(p(\theta)) + \dot{h}(p(\theta))^\top(\overline{X}_n - p(\theta)) + o(|\overline{X}_n - p(\theta)|).$$

Note that $Z_n := \sqrt{n}(\overline{X}_n - p(\theta)) = O_P(1)$, and it has an asymptotic distribution

$$Z_n \xrightarrow[n\to\infty]{d} N(0, \Sigma(\theta)),\ \Sigma(\theta) := \text{diga}(p(\theta)) - p(\theta)p(\theta)^\top,$$

and therefore we get

$$\sqrt{n}(h(\overline{X}_n) - h(p(\theta))) = \dot{h}(p(\theta))^\top Z_n + o_P(1),$$

which implies

$$\sqrt{n}(h(\overline{X}_n) - h(p(\theta))) \xrightarrow[n\to\infty]{d} N(0, \sigma_h^2(\theta)),$$

where

$$\sigma_h^2(\theta) = \dot{h}(p(\theta))^\top \Sigma(\theta) \dot{h}(p(\theta)).$$

Furthermore, we can obtain that

$$\sigma_h^2(\theta) \geq I_1^{-1}(\theta),$$

where equality holds iff

$$\dot{h}(p(\theta))^\top (X_1 - p(\theta)) = I_1^{-1}(\theta)\dot{l}_1(\theta).$$

It can be shown as following. First, note that

$$\dot{h}(p(\theta))^\top \dot{p}(\theta) = 1.$$

It implies that

$$\begin{aligned}
1 &= \dot{h}(p(\theta))^\top \frac{\partial}{\partial\theta} E_\theta X_1 \\
&= \dot{h}(p(\theta))^\top \int \frac{\partial}{\partial\theta} x f(x:\theta) d\mu(x) \\
&= \dot{h}(p(\theta))^\top \int x \frac{\partial}{\partial\theta} f(x:\theta) d\mu(x) \\
&= \dot{h}(p(\theta))^\top Cov_\theta \left(X_1, \frac{\partial}{\partial\theta} l_1(\theta) \right) \ \ (\because E_\theta \frac{\partial}{\partial\theta} f(X_1:\theta) = E_\theta \frac{\partial}{\partial\theta} l_1(\theta) = 0) \\
&= Cov_\theta \left(\dot{h}(p(\theta))^\top X_1, \frac{\partial}{\partial\theta} l_1(\theta) \right) \\
&\leq \sqrt{Var_\theta \left(\dot{h}(p(\theta))^\top X_1 \right) Var_\theta \left(\frac{\partial}{\partial\theta} l_1(\theta) \right)} \\
&= \sqrt{\dot{h}(p(\theta))^\top \Sigma(\theta) \dot{h}(p(\theta)) \cdot I_1(\theta)}
\end{aligned}$$

holds. In here, "=" holds when $\dot{h}(p(\theta))^\top X_1$ and $\partial l_1(\theta)/\partial\theta$ has a linear relationship, i.e.,

$$\dot{h}(p(\theta))^\top (X_1 - p(\theta)) = I_1^{-1}(\theta)\dot{l}_1(\theta).$$

Remark 1.3.4. Actually, previous example shows *how to deal with asymptotic distribution of FSE*, more generally.

1.3.2 Asymptotic Normality of MCE

Our real goal of this section is right here.

Theorem 1.3.5 (Asymptotic Normality of MCE)**.** *Let $X_1, \cdots, X_n$ be a random sample from*

P_θ, where $\theta \in \Theta$ and parameter space Θ is open in $\mathbb{R}^k$. Let

$$\hat{\theta}_n^{MCE} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \rho(X_i, \theta)$$

$$\theta_0 = \arg\min_{\theta\in\Theta} E_{\theta_0}\rho(X_1, \theta).$$

Under the assumption of their existence, let

$$\Psi_1(\theta) = \Psi(X_1, \theta) = \frac{\partial}{\partial\theta}\rho(X_1, \theta), \qquad \overline{\Psi}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \Psi(X_i, \theta)$$

$$\dot{\Psi}_1(\theta) = \frac{\partial \Psi_1(\theta)}{\partial\theta}, \qquad \dot{\overline{\Psi}}_n(\theta) = \frac{\partial \overline{\Psi}_n(\theta)}{\partial\theta}$$

$$\ddot{\Psi}_1(\theta) = \frac{\partial^2 \Psi_1(\theta)}{\partial\theta^2}, \qquad \ddot{\overline{\Psi}}_n(\theta) = \frac{\partial^2 \overline{\Psi}_n(\theta)}{\partial\theta^2}$$

Assume that

(A0) $P_{\theta_0}\left(\overline{\Psi}_n(\hat{\theta}_n^{MCE}) = 0\right) \xrightarrow[n\to\infty]{} 1.$

(A1) $E_{\theta_0}\Psi_1(\theta_0) = 0.$

(A2) $Var_{\theta_0}(\Psi_1(\theta_0))$ *exists.*

(A3) $E_{\theta_0}(\dot{\Psi}_1(\theta_0))$ *exists and is nonsingular.*

(A4) $\exists\delta > 0$ *and* $\exists M(X_1) = M_{\theta_0,\delta}(X_1)$ *s.t.*

$$\max_{\substack{|\theta-\theta_0|\le\delta \\ \theta\in\Theta}} |\ddot{\Psi}_1(\theta)| \le M(X_1), \text{ where } E_{\theta_0}M(X_1) < \infty.$$

(A5) $\hat{\theta}_n^{MCE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$ *as* $n \to \infty$.

Then

$$\hat{\theta}_n^{MCE} = \theta_0 + \left[-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right]^{-1}\overline{\Psi}_n(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}),$$

so that

$$\sqrt{n}(\hat{\theta}_n^{MCE} - \theta_0) \xrightarrow[n\to\infty]{d} N\left(0, \left[-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right]^{-1} Var_{\theta_0}(\Psi_1(\theta_0)) \left[-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right]^{-1}\right) \text{ under } P_{\theta_0}.$$

Remark 1.3.6 (Gradient of vector map). Let $F : \mathbb{R}^k \to \mathbb{R}^d$ be a smooth function, where

$$F(x) = (F_1(x), \cdots, F_d(x))^\top.$$

(1) (1st-order gradient)

$$\frac{\partial F}{\partial x}(x_0) := DF(x_0) = \left(\frac{\partial F}{\partial x_1}(x_0), \frac{\partial F}{\partial x_2}(x_0), \cdots, \frac{\partial F}{\partial x_k}(x_0)\right) \in \mathbb{R}^{d\times k},$$

where $\frac{\partial F}{\partial x_j}(x_0) = \left(\frac{\partial F_1}{\partial x_j}(x_0), \cdots, \frac{\partial F_d}{\partial x_j}(x_0)\right)^\top$ is a column vector. It can be interpreted as "a linear map."

(2) (2nd-order gradient)

$$\frac{\partial^2 F}{\partial x^2}(x_0) := D^2F(x_0) = \left(\frac{\partial}{\partial x_1}\frac{\partial F}{\partial x}(x_0), \cdots, \frac{\partial}{\partial x_k}\frac{\partial F}{\partial x}(x_0)\right) \in \mathbb{R}^{d\times k\times k},$$

where $\frac{\partial}{\partial x_i}\frac{\partial F}{\partial x}(x_0)$ is $d \times k$ matrix with $\frac{\partial^2 F}{\partial x_i x_j}(x_0)$ as the jth column vector. Note that it can be interpreted as "a bi-linear map."

(3) (Taylor expansion of vector-valued map)

$$\begin{aligned} F(x) &\approx F(x_0) + \sum_{j=1}^k \frac{\partial F}{\partial x_j}(x_0)(x_j - x_{0j}) + \frac{1}{2}\sum_{i,j}\frac{\partial^2 F}{\partial x_i \partial x_j}(x_0)(x_i - x_{0i})(x_j - x_{0j}) \\ &= F(x_0) + \underbrace{\frac{\partial F}{\partial x_j}(x_0)}_{\text{matrix}}\underbrace{(x - x_0)}_{\text{vector}} + \frac{1}{2}(x - x_0)^\top \underbrace{\frac{\partial^2 F}{\partial x_i \partial x_j}(x_0)}_{\text{3-array}}(x - x_0). \end{aligned}$$

In here, "matrix $DF(x_0)$ $\times$ vector" becomes a vector, and "quadratic form with 3-array $D^2F(x_0)$" becomes vector-valued. In this view, $DF(x_0)$ and $D^2F(x_0)$ can be interpreted as a linear and bi-linear map, respectively.

Proof. By Taylor's theorem, $\exists \theta_n^*$ in line$(\theta_0, \hat\theta_n)$ such that $\|\theta_n^* - \theta_0\| \le \|\hat\theta_n - \theta_0\|$ and

$$\overline{\Psi}_n(\hat\theta_n) = \overline{\Psi}_n(\theta_0) + \dot{\overline{\Psi}}_n(\theta_0)(\hat\theta_n - \theta_0) + \frac{1}{2}(\hat\theta_n - \theta_0)^\top \ddot{\overline{\Psi}}_n(\theta_n^*)(\hat\theta_n - \theta_0).$$

Also note that, from (A4) and (A5),

$$\lim_{K\to\infty}\sup_n P_{\theta_0}\left(\ddot{\overline{\Psi}}_n(\theta_n^*) \ge K\right) = 0$$

holds, and hence

$$\ddot{\overline{\Psi}}_n(\theta_n^*) = O_{P_{\theta_0}}(1),$$

i.e.,

$$\overline{\ddot{\Psi}}_n(\theta_n^*)(\hat{\theta}_n - \theta_0) = o_{P_{\theta_0}}(1).$$

($\because$ From

$$\begin{aligned}
P_{\theta_0}\left(\left|\overline{\ddot{\Psi}}_n(\theta_n^*)\right| > K\right) &= P_{\theta_0}\left(\left|\overline{\ddot{\Psi}}_n(\theta_n^*)\right| > K,\ |\hat{\theta}_n - \theta_0| \le \delta\right) + P_{\theta_0}\left(\left|\overline{\ddot{\Psi}}_n(\theta_n^*)\right| > K,\ |\hat{\theta}_n - \theta_0| > \delta\right) \\
&\le P_{\theta_0}\left(\left|\overline{\ddot{\Psi}}_n(\theta_n^*)\right| > K,\ |\hat{\theta}_n - \theta_0| \le \delta\right) + P_{\theta_0}\left(|\hat{\theta}_n - \theta_0| > \delta\right) \\
&\underset{\text{(A4)}}{\le} P_{\theta_0}\left(M(X_1) > K\right) + \underbrace{P_{\theta_0}\left(|\hat{\theta}_n - \theta_0| > \delta\right)}_{\xrightarrow[n\to\infty]{\text{(A5)}} 0},
\end{aligned}$$

for any $\epsilon > 0$ we get for large N

$$\sup_{n>N} P_{\theta_0}\left(|\overline{\ddot{\Psi}}_n(\theta_n^*)| > K\right) \le \frac{1}{K} E_{\theta_0} M(X_1) + \epsilon,$$

which implies

$$\lim_{K\to\infty} \sup_{n>N} P_{\theta_0}\left(|\overline{\ddot{\Psi}}_n(\theta_n^*)| > K\right) = 0.$$

(Let N more large and take $\epsilon \searrow 0$)) Thus, we get

$$\begin{aligned}
\overline{\Psi}_n(\hat{\theta}_n) &= \overline{\Psi}_n(\theta_0) + \overline{\dot{\Psi}}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \overline{\ddot{\Psi}}_n(\theta_n^*)(\hat{\theta}_n - \theta_0) \\
&= \overline{\Psi}_n(\theta_0) + \left(\overline{\dot{\Psi}}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \overline{\ddot{\Psi}}_n(\theta_n^*)\right)(\hat{\theta}_n - \theta_0) \\
&= \overline{\Psi}_n(\theta_0) + \left(\overline{\dot{\Psi}}_n(\theta_0) + o_{P_{\theta_0}}(1)\right)(\hat{\theta}_n - \theta_0) \\
&= \overline{\Psi}_n(\theta_0) + \left(E_{\theta_0}\dot{\Psi}_1(\theta_0) + o_{P_{\theta_0}}(1)\right)(\hat{\theta}_n - \theta_0).
\end{aligned}$$

Now, note that

(1) $P_{\theta_0}(\overline{\Psi}_n(\hat{\theta}_n) = 0) \xrightarrow[n\to\infty]{} 1$.

(2) $E_{\theta_0}\dot{\Psi}_1(\theta_0) + o_{P_{\theta_0}}(1)$ is nonsingular with probability 1 (See remark 1.3.7)

(3) If $X_n = Y_n + o_P(a_n)$ on $\mathcal{E}_n$ with $P(\mathcal{E}_n) \xrightarrow[n\to\infty]{} 1$, then $X_n = Y_n + O_P(a_n)$ (on whole space), and the same holds for o_P (See remark 1.2.7).

Thus, on the set with probability tending to 1,

$$\hat{\theta}_n - \theta_0 = \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0) + o_{P_{\theta_0}}(1)\right)^{-1} \overline{\Psi}_n(\theta_0)$$

$$= \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1}\overline{\Psi}_n(\theta_0) + o_{P_{\theta_0}}(1)$$

holds, which yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1}\sqrt{n}\overline{\Psi}_n(\theta_0) + o_{P_{\theta_0}}(1)$$

and therefore

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{d} N\left(0, \left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1} Var_{\theta_0}\Psi_1(\theta_0)\left(-E_{\theta_0}\dot{\Psi}_1(\theta_0)\right)^{-1}\right).$$

□

Remark 1.3.7. If $A \in \mathbb{R}^{d\times d}$ is symmetric positive definite matrix, then for small perbutation Δ s.t. $\|\Delta\|_2 < \sigma_{min}(A)$, $\text{rank}(A+\Delta) = d$. Note that $\text{rank}(A) = d$. In here, $\sigma_{min}(A)$ denotes the smallest eigenvalue of A, and $\|\cdot\|_p$ is a matrix norm induced by corresponding $\mathcal{L}^p$ vector norm, i.e.,

$$\|\Delta\|_p = \sup_{x:\|x\|_p=1}\|\Delta x\|_p.$$

Proof. (Motivation: for $c \approx 0$ and $x \neq 0$,

$$\frac{1}{x+c} = \frac{x}{c} - \frac{x^2}{c^2} + \frac{x^3}{c^3} - \cdots$$

exists, or for small vectors u, v, $A + uv^\top$ is nonsingular from

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + u^\top A^{-1}v},$$

if A is invertible) Let $\text{rank}(A+\Delta) < d$. Then $\exists x_0 \neq 0$ s.t. $(A+\Delta)x_0 = 0$ and $\|x_0\|_2 = 1$. Then by definition of matrix norm,

$$\|\Delta\|_2 \geq \|\Delta x_0\|_2 = \|Ax_0\|_2 \geq \sigma_{min}(A)$$

holds. The last inequality is from spectral theorem. It is contradictory to our assumption that Δ is small. □

1.3.3 Asymptotic Normality and Efficiency of MLE

Note that MLE is just a special case of MCE.

Theorem 1.3.8. *Let $X_1, \cdots, X_n$ be a random sample from P_θ, where $\theta \in \Theta$ and parameter space Θ is open in $\mathbb{R}^k$. Recall that MLE is an MCE with*

$$\rho(x, \theta) = -\log p_\theta(x), \ p_\theta : \text{pdf of } P_\theta.$$

Under the assumption of their existence, denote

$$\begin{aligned} \dot{l}_n(\theta) &= \sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(X_i), & \bar{\dot{l}}_n(\theta) &= \frac{1}{n}\dot{l}_n(\theta) \\ \ddot{l}_n(\theta) &= \frac{\partial^2}{\partial\theta^2} \log p_\theta(X_i), & \bar{\ddot{l}}_n(\theta) &= \frac{1}{n}\ddot{l}_n(\theta) \\ \dddot{l}_n(\theta) &= \frac{\partial^3}{\partial\theta^3} \log p_\theta(X_i), & \bar{\dddot{l}}_n(\theta) &= \frac{1}{n}\dddot{l}_n(\theta). \end{aligned}$$

Also assume that

(M0) $P_{\theta_0}\left(\dot{l}_n(\hat{\theta}_n^{MLE}) = 0\right) \xrightarrow[n\to\infty]{} 1.$

(M1) $E_{\theta_0}\dot{l}_1(\theta_0) = 0.$

(M2) $I(\theta_0) = Var_{\theta_0}(\dot{l}_1(\theta_0))$ *exists.*

(M3) $E_{\theta_0}(\ddot{l}_1(\theta_0))$ *exists and is nonsingular.*

(M4) $\exists\delta_0 > 0$ *and* $\exists M(X_1) = M_{\theta_0,\delta}(X_1)$ *s.t.*

$$\max_{\substack{|\theta-\theta_0|\le\delta_0 \\ \theta\in\Theta}} |\dddot{l}_1(\theta)| \le M(X_1), \ \text{where } E_{\theta_0}M(X_1) < \infty.$$

(M5) $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$ *as* $n \to \infty$.

(M6) $I(\theta_0) = E_{\theta_0}(-\ddot{l}_1(\theta_0)).$

Under (M0) $\sim$ (M6),

$$\hat{\theta}_n^{MLE} = \theta_0 + I(\theta_0)^{-1}\bar{\dot{l}}_n(\theta_0) + o_{P_{\theta_0}}(n^{-1/2}),$$

so that

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow[n\to\infty]{d} N\left(0, I(\theta_0)^{-1}\right).$$

Even though MCE is more general version of MLE, we frequently use MLE to estimate the parameter. Following theorem says that, "MLE is more (asymptotically) efficient than MCE," i.e., log contrast function makes MCE the most efficient.

Theorem 1.3.9 (Asymptotic efficiency of MLE). *Assume (A0) $\sim$ (A6), and (M0) $\sim$ (M6) hold, where*

$$\text{(A6)}: E_{\theta_0}\dot{\Psi}_1(\theta_0) = -E_{\theta_0}\dot{l}_1(\theta_0)\Psi_1^\top(\theta_0).$$

Then

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow[n\to\infty]{d} N(0, I(\theta_0)^{-1})$$

$$\sqrt{n}(\hat{\theta}_n^{MCE} - \theta_0) \xrightarrow[n\to\infty]{d} N(0, \Sigma_\Psi(\theta_0))$$

with $\Sigma_\Psi(\theta_0) - I(\theta_0)^{-1}$ being nonnegative definite (i.e., "$\Sigma_\Psi(\theta_0) \ge I(\theta_0)^{-1}$"), and "=" holds if and only if

$$\Psi_1(\theta_0) = (-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1(\theta_0)^{-1}\dot{l}_1(\theta_0). \qquad \text{("Determining contrast function")}$$

Proof. By Cauchy-Schwarz inequality,

$$\begin{aligned}
Corr(\lambda^\top \dot{l}_1(\theta_0), \gamma^\top\Psi_1(\theta_0)) &= \frac{\lambda^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}}\\
&= \frac{\lambda^\top I_1^{1/2}(\theta_0)I_1^{-1/2}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}}\\
&\le \frac{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}}{\sqrt{\lambda^\top I_1(\theta_0)\lambda}\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}}\\
&= \frac{\sqrt{\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}}{\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}}
\end{aligned}$$

holds, and hence

$$\max_{\lambda\neq 0} Corr(\lambda^\top \dot{l}_1(\theta_0), \gamma^\top\Psi_1(\theta_0)) = \frac{\sqrt{\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma}}{\sqrt{\gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma}}$$

is obtained (we can find λ that achieves maximum). Since correlation coefficient is less than 1, we get

$$\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))I_1^{-1}(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))\gamma \le \gamma^\top Var_{\theta_0}\Psi_1(\theta_0)\gamma$$

for any γ. Using $(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}\gamma$ instead of γ, we obtain that

$$\gamma^\top I_1^{-1}(\theta_0)\gamma \le \gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}Var_{\theta_0}\Psi_1(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}\gamma,$$

i.e.,

$$I_1^{-1}(\theta_0) \leq (-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}Var_{\theta_0}\Psi_1(\theta_0)(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1} = \Sigma_\Psi(\theta_0).$$

Equality holds iff

$$\gamma^\top(-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}I_1^{-1}(\theta_0)\dot{l}_1(\theta_0) = \gamma^\top\Psi_1(\theta_0),$$

i.e.,

$$\Psi_1(\theta_0) = (-E_{\theta_0}\dot{\Psi}_1(\theta_0))^{-1}I_1^{-1}(\theta_0)\dot{l}_1(\theta_0).$$

□

Remark 1.3.10. The claim that the MLE has smaller variance than other asymptotically normal estimators was known as *Fisher's conjecture*. This is true for a certain class of estimators in a *regular* parametric model. Essential property for such a comparison is the "*uniform*" convergence to the normal distribution as it can be seen in the following example.