STAT5005

Measure Theory

Sec1. Probability Space

Def. σ-field $\mathcal{F}$: i) $\Omega \in \mathcal{F}$; ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$; iii) if $A_1, A_2, \dots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
algebra: iii) if $A_1, \dots A_n \in \mathcal{F}$, then $\cup_{i=1}^{n} A_i \in \mathcal{F}$.

Fact. if $\mathcal{F}_i, i \in I$ are all σ-fields, $\cap_{i \in I}\mathcal{F}_i$ is a σ-field.

Def. measure μ: i) $\mu(A) \geq 0, \forall A \in \mathcal{F}$; ii) $\mu(\emptyset) = 0$; iii) if $A_1, A_2, \dots \in \mathcal{F}$ disjoint, then $\mu(\cup_i A_i) = \sum_i \mu(A_i)$.

Thm 1.1.1. iii) continuity from below: if $A_i \uparrow A$, then $\mu(A_i) \uparrow \mu(A)$. iv) continuity from above: if $A_i \downarrow A$ and $\mu(A_1) < \infty$, then $\mu(A_i) \downarrow \mu(A)$.

Def. Borel σ-field: $\mathcal{R} = \sigma(\{(a,b] : -\infty < a < b < \infty\})$.

Def. π-system $\mathcal{P}$: if $A, B \in \mathcal{P}$, then $A \cap B \in \mathcal{P}$.
λ-system $\mathcal{L}$: i) $\Omega \in \mathcal{L}$; ii) if $A, B \in \mathcal{L}$ and $A \subset B$, then $B \backslash A \in \mathcal{L}$; iii) if $A_1, A_2, \dots \in \mathcal{L}$, and $A_i \uparrow A$, then $A \in \mathcal{L}$.

Ex $\star$. if $\mathcal{F}$ is π-system and λ-system, then $\mathcal{F}$ is σ-field.

Thm 2.1.2. if $\mathcal{P}$ is π-system, $\mathcal{L}$ is λ-system, $\mathcal{P} \subset \mathcal{L}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.

Sec2-1. Measurable Function

Def. f is meas- if $\{\omega_1 \in \Omega_1 : f(\omega_1) \in A\} = f^{-1}(A) \in \mathcal{F}_1, \forall A \in \mathcal{F}_2$.

Fact. gen- σ-field by f: $\sigma(f) = \{f^{-1}(A) : A \in \mathcal{F}_2\}$ is σ-field in Ω_1, $\{A \subset \Omega_2 : f^{-1}(A) \in \mathcal{F}_1\}$ is a σ-field in Ω_2.

Thm 1.3.1. if $\mathcal{F}_2 = \sigma(\mathcal{A}_2)$, and $f^{-1}(\mathcal{A}_2) \subset \mathcal{F}_1$, then f is meas-.

Thm 1.3.2. if f_1 and f_2 are meas-, then $f_2 \circ f_1$ is meas-.

Def. induced measure: $\mu_2(A) = \mu_1(f^{-1}(A))$.

Sec2-2. Random Variable

Thm 1.3.5. $\inf_n X_n, \sup_n X_n, \limsup X_n, \liminf X_n$ are r-v-.

Ex 1.3.1. $\sigma(X^{-1}(\mathcal{A})) = X^{-1}(\sigma(\mathcal{A}))$
Hint: $\mathcal{C} = \{B \in \sigma(\mathcal{A}) : X^{-1}(B) \in \sigma(X^{-1}(\mathcal{A}))\}$.

Sec2-3. Distribution

Thm 1.2.2. $\Omega = (0,1), \mathcal{F} = \mathcal{R}, P =$ Lebesgue measure, then $X(\omega) = F^{-1}(\omega) = \inf\{y \in \mathbb{R} : F(y) \geq \omega\} = \sup\{y \in \mathbb{R} : F(y) < \omega\}$ with dist- F.
Hint: $\{\omega : \omega \leq F(x)\} = \{\omega : X(\omega) \leq x\}$, right-continuous of F.

Sec3. Expectation

Def. indicator -> simple -> non-negative -> arbitrary.
case3: $\mathbb{E}[X] = \sup\{\mathbb{E}[Y] : 0 \leq Y \leq X, Y \text{ is simple}\}$.

Prop. a) monotonicity, b) linearity.
Hint: $Z_M^{(u)} = \frac{1}{2^M}\lceil 2^M Z\rceil, Z_M^{(l)} = \frac{1}{2^M}\lceil 2^M Z - 1\rceil$ for truc- case.

Thm MC. X_n n-n seq- r-v-. if $X_n \uparrow X$, then $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$.
Hint: $Y_\epsilon = \sum_i (b_i - \epsilon/2)1_{B_i}$.

Thm Fatou's L. if $X_n \geq 0$, then $\liminf_n \mathbb{E}[X_n] \geq \mathbb{E}[\liminf_n X_n]$.

Thm DC. if $X_n \to X$, $|X_n| \leq Y$ with $\mathbb{E}[Y] < \infty$, then $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$.

Thm Jensen. if $\mathbb{E}[X] < \infty$, φ is convex, then $\mathbb{E}[|\varphi(X)|] \leq \infty$.

Thm Hölder. if $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then $\mathbb{E}[|XY|] \leq \|X\|_p + \|Y\|$
Hint: $x \cdot y \leq x^p/p + x^q/q$ via concav- of log.

Thm Minkowski. for $p \geq 1$, $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$.

Thm Markov. if r-v- $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{1}{a}\mathbb{E}[X]$.

Thm Chebyshev. if $\exists$ var, then $P(|X - \mathbb{E}[X]| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(X)$.

Law of Large Number

Sec1. Independence

Def. inde- events -> collections (σ-fields) -> random variables.

Thm 2.1.3. if π-sys- $\mathcal{A}_i|_{i=1}^n$ are inde-, then $\sigma(\mathcal{A}_i)|_{i=1}^n$ are indep-

Proof. It suffices to show that

$$P(B_1 \cdots B_n) = P(B_1)\cdots P(B_n), \quad \forall B_i \in \sigma(\mathcal{A}_i).$$

Fix $B_i \in \mathcal{A}_i$ or $B_i = \Omega$ for $i = 2, \dots, n$. Define

$$\mathcal{L}_1 := \{B_1 \in \sigma(\mathcal{A}_1) : P(B_1 \cdots B_n) = P(B_1)\cdots P(B_n)\}.$$

Note that $\mathcal{L}_1$ contains the π-system $\mathcal{A}_1$ by assumption. Moreover, it can be easily verified that $\mathcal{L}_1$ is a λ-system (details omitted). Therefore, by the π-λ theorem, $\sigma(\mathcal{A}_1) \subset \mathcal{L}_1$. We have proved that $\sigma(\mathcal{A}_1), \mathcal{A}_2, \dots, \mathcal{A}_n$ are independent. Repeat the above argument to $\mathcal{A}_2, \dots \mathcal{A}_n$

Thm 2.1.4. r-v- $X_i|_{i=1}^n$ are inde- if-f- $P(\cap_i\{X_i \leq x_i\}) = \prod_i P(X_i \leq x_i)$.

Thm -. if $X_i|_{i=1}^n$ are inde-, then $\sigma(X_i : i \in I) \perp\!\!\!\perp \sigma(X_j : j \in I^c)$.

Thm 2.1.5. if $X_i|_{i=1}^n$ inde-, then $g(X_i, i \in I) \perp\!\!\!\perp h(X_j, j \in I^c)$, g, h meas-.

Thm 2.1.8. if X, Y inde-, $\mathbb{E}[\cdot] < \infty$ or ≥ 0, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

$$E(XY) = \sum_{i=1}^n\sum_{j=1}^m a_i b_j P(A_i B_j) = \sum_{i=1}^n\sum_{j=1}^m a_i b_j P(A_i)P(B_j) = E(X)E(Y).$$

Case 3. $X, Y \geq 0$. Choose a sequence of simple random variables $X_n \uparrow X$ and $\sigma(X_n) \subset \sigma(X)$. ($X_M$ can be chosen as $X_M^{(l)}$ as in Chapter 1, Section 3.2.) Choose also a sequence of simple random variables $Y_n \uparrow X$ and $\sigma(Y_n) \subset \sigma(X)$. From Case 2, we have $E(X_nY_n) = E(X_n)E(Y_n)$. Then take the limit as $n \to \infty$ to obtain (1.1).

Thm Kolmogorov's 0-1 Law. if X_i's inde-, tail σ-field $\mathcal{T} = \cap_n \sigma(X_k, k \geq n)$, then $P(A) = 0, 1$ for $A \in \mathcal{T}$.

Proof. Step 1: $\sigma(X_{k+1}, X_{k+2}, \dots)$ can be written as $\sigma(\cup_{j=1}^{\infty}\sigma(X_{k+1}, \dots, X_{k+j}))$. The union inside of $\sigma(\dots)$ is a π-system. Moreover, $\sigma(X_1, \dots, X_k)$ is independent of this π-system by the condition. Therefore, $\sigma(X_1, \dots, X_k)$ is independent of $\sigma(X_{k+1}, X_{k+2}, \dots)$.

Step 2: Write $\sigma(X_1, X_2, \dots) = \sigma(\cup_{j=1}^{\infty}\sigma(X_1, \dots, X_j))$ and argue similarly that $\sigma(X_1, X_2, \dots)$ is independent of $\mathcal{T}$.

Step 3: By definition, $\mathcal{T} \subset \sigma(X_1, X_2, \dots)$.

Therefore $\mathcal{T}$ is independent of itself. The theorem then follows from the previous Fact. □

Sec2-1. Weak Law of Large Number

Def. converges in prob: if $\forall\ \epsilon > 0$, $P(|Y_n - Y| > \epsilon) \to 0$, as $n \to \infty$.
converges in L_p $E|Y_n - Y|^p \to 0$, as $n \to \infty$. $P(|Y_n - Y| > \epsilon) \leq \frac{E|Y_n - Y|^p}{\epsilon^p} \to 0$

Theorem. If $(S_n)_{n=1}^{\infty}$ is a sequence of random variables with $\sigma_n^2 = \mathrm{Var}(S_n)$ and $\sigma_n^2/b_n^2 \to 0$, then

$$\frac{S_n - E(S_n)}{b_n} \to 0 \quad \text{in } L_2 \text{ and in probability.}$$

Example 2.2.3 Coupon Collector's Problem. Let $n \geq 1$ be an integer and $X_1, X_2, \dots$ be independent and uniformly distributed on $\{1, \dots, n\}$. Define for $k = 0, 1, \dots, n$

$$\tau_k^n = \inf\{m \geq 0 : |\{X_1, \dots, X_m\}| = k\},$$

where $|\cdot|$ denotes the size of a set (the number of distinct elements in a set). Let $T_n = \tau_n^n$. Then we have

$$\frac{T_n}{n\log n} \to 1 \quad \text{in probability.}$$

Proof. Let $Y_k = \tau_k^n - \tau_{k-1}^n$, $k = 1, 2, \dots, n$. Then Y_k follows a geometric distribution with parameter $\frac{n-k+1}{n}$ and

BC Proof: (i) $P(A_n \text{ i.o.}) = P(\bigcap \bigcup A_k) = \lim P(\bigcup A_k) = 0$
(ii) $P(A_n \text{ i.o.}) = 1 - P(\bigcup \bigcap A_k^c) = 1 - \lim P(\bigcap A_k^c) = 1 - \lim \prod P(A_k^c)$
$\geq 1 - \lim \prod e^{-P(A_k)} = 1$

$$EY_k = \frac{n}{n-k+1}, \quad \mathrm{Var}(Y_k) = \frac{k-1}{n}\frac{n^2}{(n-k+1)^2}.$$

Moreover, $Y_1, \dots, Y_n$ are independent and $T_n = \sum_{k=1}^n Y_k$. Therefore,

$$E(T_n) = n(1 + \frac{1}{2} + \dots + \frac{1}{n}) \sim n\log n$$

and

$$\mathrm{Var}(T_n) = \sum_{k=1}^n \mathrm{Var}(Y_k) \leq \sum_{k=1}^n \frac{n^2}{(n-k+1)^2} \leq Cn^2$$

for some absolute constant C. Choosing $b_n = n\log n$ in the above theorem, we obtain

$$\frac{T_n - n\log n}{n\log n} \to 0 \quad \text{in probability.}$$

THM2.2.7 $X_{nk} = X_k$ $b_n = n$
$\bar{X}_{nk} = X_k 1\{|X_k| < n\}$
(c1) $= \sum_{k=1}^n P(|X_k| > n) = nP(|X_1| > n) \to 0$
(c2) $= n^{-2}\sum_k E[(X_k 1\{|X_k| < n\})^2]$
$= n^{-1}E[\bar{X}_1^2 1\{\ \}] = n^{-1}\int_0^n 2yP(|X_1| > y)dy$

Thm 2.2.6. (WLLN for triangular arrays)
suppose that $\{X_{n,k} : 1 \leq k \leq n\}$ are inde-, $\bar{X}_{n,k} = X_{n,k}1_{\{|X_{n,k}| \leq b_n\}}$,
if i) $\sum_{k=1}^n P(|X_{n,k}| \geq b_n) \to 0$, ii) $b_n^{-2}\sum_{k=1}^n \mathbb{E}[(\bar{X}_{n,k})^2] \to 0$, (ii) Cheby
$a_n = \sum_{k=1}^n \mathbb{E}[\bar{X}_{n,k}]$ then $(S_n - a_n)/b_n \to 0$ in prob- $P(|\ X_{nk}| \leq (i) + P(|\frac{S_n - a_n}{b_n}| > \epsilon)$

Thm 2.2.7. (WLLN without moment assumption) $\mu_n = \mathbb{E}[X1_{\{X \leq n\}}]$.
i.i.d. if $xP(|X_1| > x) \to 0$ as $x \to \infty$, then $\frac{1}{n}S_n - \mu_n \to 0$ in prob-.
Remark: for $0 < \epsilon < 1$, $\mathbb{E}[|X|^{1-\epsilon}] < \infty$. 2.2.9 By DCT, $xP(|X_1| > x) \leq E|X_1|1\{|X_1| > x\} \to 0$ DCT, $EX_1 1\{|X_1| \leq n\} \to \mu$

Lem 2.2.8. if $Y \geq 0$ and $p > 0$, then $\mathbb{E}[Y^p] = \int_0^{\infty} py^{p-1}P(Y > y)dy$.

Thm 2.2.9. (WLLN with finite 1st moment)
i.i.d. $\mu = \mathbb{E}[X_1]$. if $\mathbb{E}[|X_1|] < \infty$, then $\frac{1}{n}S_n - \mu \to 0$ in prob.

Sec2-2. Strong Law of Large Number

Def. events A_n occurs infinitely often $\{A_n \text{ i.o.}\} = \cap_n \cup_k A_k$.
$= \limsup A_n = \{w : \exists$ a subsequence n_k, $w \in \cap_k A_{n_k}\}$.

Fact. $Y_n \to Y$ a.s. if-f- $\forall \epsilon > 0, P(|Y_n - Y| > \epsilon \text{ i.o.}) = 0$

$$P(\limsup_{n\to\infty}|Y_n - Y| = 0) = \lim_{\varepsilon \downarrow 0} P(\limsup_{n\to\infty}|Y_n - Y| \leq \varepsilon) = 1.$$

Thm 2.3.1/6. (Borel-Cantelli Lemma)
i) if $\sum_n P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.
ii) if $\sum_n P(A_n) = \infty$ and A_n's indep-, then $P(A_n \text{ i.o.}) = 1$.

Theorem. If $Y_n \to Y$ a.s., then $Y_n \to Y$ in probability. If $Y_n \to Y$ in probability, then there exists a subsequence n_k such that $Y_{n_k} \to Y$ a.s.

Proof. The first statement follows directly by definition. To prove the second statement, we choose a subsequence $n_1 < n_2 < n_3 < \dots$ such that

$$P(|Y_{n_k} - Y| > \frac{1}{2^k}) \leq \frac{1}{2^k}.$$

2.5.2 $P\{\ \} = \sum_{k=1}^n P(A_k) \leq \sum_{k=1}^n \frac{ES_k^2 1_A}{x^2}$
$ES_n^2 1_{A_k} = E[(S_k + S_n - S_k)^2 1_{A_k}] \geq ES_k^2 1_A$
Since $E[S_k 1_A(S_n - S_k)] = 0$ (indep)

Such a subsequence exists because of the condition that $Y_n \to Y$ in probability. For this subsequence, we have

$$\sum_{k=1}^{\infty} P(|Y_{n_k} - Y| > \frac{1}{2^k}) < \infty.$$

By B-C lemma (i),

$$P(|Y_{n_k} - Y| > \frac{1}{2^k} \ i.o.) = 0.$$

The implies $P(|Y_{n_k} - Y| > \varepsilon \ i.o.) = 0$ for any $\varepsilon > 0$; hence almost sure convergence. □

Theorem 2.3.8. If $A_1, A_2, \dots$ are pairwise independent, and $\sum_{i=1}^n P(A_i) = \infty$. Then, as $n \to \infty$,

$$\frac{\sum_{i=1}^n 1_{A_i}}{\sum_{i=1}^n P(A_i)} \to 1 \quad a.s.$$

Proof. Let $S_n = \sum_{i=1}^n 1_{A_i}, b_n = \sum_{i=1}^n P(A_i)$. We have, for $\varepsilon > 0$

$$P(|\frac{S_n}{b_n} - 1| > \varepsilon) \leq \frac{\mathrm{Var}(S_n/b_n)}{\varepsilon^2} = \frac{\sum_{i=1}^n P(A_i)(1 - P(A_i))}{\varepsilon^2 b_n^2}$$
$$\leq \frac{\sum_{i=1}^n P(A_i)}{\varepsilon^2(\sum_{i=1}^n P(A_i))^2} = \frac{1}{\varepsilon b_n}.$$

Choose a subsequence $n_k = \min\{m : b_m \geq k^2\}$ (it is possible because of the condition $\sum_{i=1}^n P(A_i) = \infty$). Then

$$\sum_{k=1}^{\infty} P(|\frac{S_{n_k}}{b_{n_k}} - 1| > \varepsilon) < \infty,$$

and by B-C lemma (i),

$$\frac{S_{n_k}}{b_{n_k}} \to 1 \quad a.s.$$

For any sufficiently large integer m, there exists k such that $n_{k-1} \leq m \leq n_k$, and

$$\frac{S_{n_{k-1}}}{b_{n_{k-1}}}\frac{b_{n_{k-1}}}{b_{n_k}} = \frac{S_{n_{k-1}}}{b_{n_k}} \leq \frac{S_m}{b_m} \leq \frac{S_{n_k}}{b_{n_{k-1}}} = \frac{S_{n_k}}{b_{n_k}}\frac{b_{n_k}}{b_{n_{k-1}}}.$$

Because both sides tend to 1 a.s., S_m/m in the middle must tend to 1 a.s.. □

Theorem 2.3.5. Let $X_1, X_2, \dots$ be i.i.d. with $EX_i = \mu, E(X_i^4) < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \to \mu \quad a.s.$$

Proof. WLOG, assume $\mu = 0$ (otherwise, consider $X_i' = X_i - \mu$). We have

$$P(|\frac{S_n}{n}| > \varepsilon) \leq \frac{E[S_n^4]}{\varepsilon^4 n^4} \leq \frac{\sum_{i,j=1}^n E[X_i^2X_j^2]}{\varepsilon^4 n^4} \leq \frac{C}{\varepsilon^4 n^2},$$

Thm SLLN. i.i.d. $\mathbb{E}[|X_i|] < \infty$, then $S_n/n \to \mu$ a.s.

[Step 1: Truncation.] Let $Y_k = X_k 1_{\{|X_k| \leq k\}}$, $T_n = \sum_{i=1}^n Y_i$. Because

$$\sum_{k=1}^{\infty} P(|X_k| > k) \leq E|X_1| < \infty,$$

we have by B-C lemma (i)

$$P(X_k \neq Y_k \ i.o.) = 0.$$

Therefore, to prove the theorem, if suffices to show $\frac{T_n}{n} \to \mu$ a.s..

[Step 2: 2nd moment calculation.] Fix $\alpha > 1$. Choose $k(n) = \lfloor \alpha^n \rfloor$ (this means the integer part, and it is obviously $\geq \alpha^n/2$). We have, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left(|\frac{T_{k(n)} - ET_{k(n)}}{k(n)}| > \varepsilon\right) \leq 4(1 - \alpha^{-2})^{-1}\varepsilon^{-2}\sum_{m=1}^{\infty}\frac{\mathrm{Var}(Y_m)}{m^2}.$$

$$\sum_{m=1}^{\infty}\frac{\mathrm{Var}(Y_m)}{m^2} \leq \sum_{m=1}^{\infty}\frac{EY_m^2}{m^2} = \sum_{m=1}^{\infty}\frac{\int_0^m 2yP(|Y_m| > y)dy}{m^2} \leq C\int_0^{\infty}P(|X_1| > y)dy = CE|X_1| < \infty$$

$$\sum_{n=1}^{\infty} P\left(|\frac{T_{k(n)} - ET_{k(n)}}{k(n)}| > \varepsilon\right) < \infty,$$

and by B-C lemma (i),

$$\frac{T_{k(n)} - ET_{k(n)}}{k(n)} \to 0 \quad a.s.$$

Note also that, because $EY_m = EX_1 1_{\{|X_1| \leq m\}} \to \mu$ as $m \to \infty$, we have

$$\frac{ET_{k(n)}}{k(n)} = \frac{\sum_{m=1}^{k(n)} EY_m}{k(n)} \to \mu.$$

Therefore,

$$\frac{T_{k(n)}}{k(n)} \to \mu \quad a.s.$$

[Step 3: Subsequence method.] For $k(n) \leq m < k(n+1)$, we have

$$\frac{k(n)}{k(n+1)}\frac{T_{k(n)}}{k(n)} \leq \frac{T_m}{m} \leq \frac{k(n+1)}{k(n)}\frac{T_{k(n+1)}}{k(n+1)}.$$

The left-hand side tends to μ/α and the right-hand side tends to $\mu\alpha$. Therefore, for any $\alpha > 1$,

$$\limsup_{m\to\infty}\frac{T_m}{m} \leq \mu\alpha, \quad \liminf_{m\to\infty}\frac{T_m}{m} \geq \mu/\alpha.$$

The theorem follows by letting $\alpha \downarrow 1$. □

Remark. Let $X_1, X_2, \dots$ be i.i.d. and $S_n = \sum_{i=1}^n X_i$. If $\frac{S_n}{n} \to \mu$ a.s. for some finite number μ, then we must have $E|X_i| < \infty$. This means the condition of the above SLLN is necessary. (Reverse of SLLN)

Proof of the Remark. Consider $\frac{X_n}{n} = \frac{S_n}{n} - \frac{S_{n-1}}{n-1}\frac{n-1}{n}$. This tends to 0 a.s. because of the condition $\frac{S_n}{n} \to \mu$ a.s.. Therefore, $P(|\frac{X_n}{n}| > 1 \ i.o.) = 0$ and by the B-C lemma (ii),

$$\sum_{n=1}^{\infty} P(|\frac{X_n}{n}| > 1) < \infty.$$

Therefore, from the i.i.d. assumption,

$$E|X_1| \leq 1 + \sum_{n=1}^{\infty} P(|X_1| > n) < \infty.$$

□

Next result concerns the case of infinite mean.

Theorem. Let $X_1, X_2, \dots$ be i.i.d. and $S_n = \sum_{i=1}^n X_i$. If $E(X_i^+) = \infty$ and $E(X_i^-) < \infty$, then

$$\frac{S_n}{n} \to \infty \quad a.s.$$

Proof. Define $X_i^M = X_i \wedge M$. Then $E|X_i^M| < \infty$. By SLLN, for $S_n^M := \sum_{i=1}^n X_i^M$, we have

$$\frac{S_n^M}{n} \to EX_i^M \quad a.s.$$

Therefore,

$$\liminf_{n\to\infty}\frac{S_n}{n} \geq \lim_{n\to\infty}\frac{S_n^M}{n} = EX_i^M \to \infty, \text{ as } M \to \infty.$$

Sec3. Convergence of Random Series

Thm 2.5.2. (Kolmogorov's Maximal Inequality)
indep-, $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] < \infty$, then $P(\max_{k \leq n}|S_k| \geq x) \leq \mathbb{E}[S_n^2]/x^2$.
Hint: $A_k = \{|S_i|$ for $i < k, |S_k| \geq x\}$.

Thm 2.5.3. indep-, $\mathbb{E}[X_i] = 0$. $\sum_i \mathbb{E}[X_i^2] < \infty$ => $\sum_i X_i$ converges a.s.
Hint: $\omega_M = \sup_{m,n > M}|S_m - S_n| \downarrow 0$ a.s. as $M \to \infty$.

Proof. Let $S_n = \sum_{i=1}^n X_i$. It suffices to prove S_n is a Cauchy sequence, a.s., that is,

$$\omega_M := \sup_{m,n \geq M}|S_m - S_n| \downarrow 0 \quad a.s. \quad \text{as } M \uparrow \infty.$$

To show this, for any $\varepsilon > 0$, we write

$$P(\omega_M > 2\varepsilon \ i.o.) = P(\omega_M > 2\varepsilon, \forall M) \quad \text{(by monotonicity of } \omega_M)$$
$$= P(\cap_M\{\omega_M > 2\varepsilon\}) = \lim_{M\to\infty} P(\omega_M > 2\varepsilon).$$

Note that, by definition of ω_M and the union bound,

$$P(\omega_M > 2\varepsilon) = P(\sup_{m,n \geq M}|S_m - S_n| > 2\varepsilon)$$
$$\leq P(\sup_{m \geq M}|S_m - S_M| > \varepsilon) + P(\sup_{n \geq M}|S_n - S_M| > \varepsilon)$$
$$= 2P(\sup_{m \geq M}|S_m - S_M| > \varepsilon).$$

Moreover, by the Kolmogorov's maximal inequality,

$$P(\sup_{m \geq M}|S_m - S_M| > \varepsilon) = \lim_{N\to\infty} P(\sup_{M \leq m \leq N}|S_m - S_M| > \varepsilon)$$
$$\leq \limsup_{N\to\infty}\frac{\mathrm{Var}(S_N - S_M)}{\varepsilon^2} = \frac{1}{\varepsilon^2}\sum_{i=M+1}^{\infty}E(X_i^2) \to 0.$$

Combining the above arguments, we conclude that

$$P(\omega_M > 2\varepsilon \ i.o.) = 0, \quad \forall\ \varepsilon > 0,$$

hence $\omega_M \to 0$. □

Example. Let $X_1, X_2, \dots$ be i.i.d. Rademacher variables, i.e., $P(X_i = 1) = P(X_i = -1) = \frac{1}{2}$. We have, for $\alpha > 1/2$, $\sum_{i=1}^{\infty}\frac{X_i}{i^\alpha}$ converges a.s.

Thm 2.5.4. (Kolmo-'s three-series thm) X_i indep-, $Y_i = X_i 1_{\{|X_i| < A\}}$, if i) $\sum^{\infty} P(|X_n| > A) < \infty$, ii) $\sum^{\infty} \mathbb{E}[Y_n]$ converges, iii) $\sum \mathrm{Var}(Y_n) < \infty$, then $\sum X_n$ converges a.s.

Proof. Let $\mu_n = EY_n$. From Theorem 2.5.3 above and condition (iii), we have $\sum_{n=1}^{\infty}(Y_n - \mu_n)$ converges a.s. Combined with condition (ii), we have $\sum_{n=1}^{\infty} Y_n$ converges a.s. From the B-C lemma and condition (i), we have

$$P(X_n \neq Y_n \ i.o.) = 0.$$

Therefore, $\sum_{n=1}^{\infty} X_n$ converges a.s. □

Theorem 2.5.5 (Kronecker's Lemma). If $a_n \uparrow \infty$ and $\sum_{n=1}^{\infty}\frac{x_n}{a_n}$ converges, then $\frac{\sum_{i=1}^n x_n}{a_n} \to 0$.

Second Proof. Recall that WLOG, we can assume $X_i \geq 0$. As in the first proof, we let $Y_k = X_k 1_{\{|X_k| \leq k\}}$, $T_n = \sum_{i=1}^n Y_i$. We have argued that $P(X_k \neq Y_k \ i.o.) = 0$ and it suffices to show $\frac{T_n}{n} \to \mu$ a.s.. Recall that we have also argued that $\frac{ET_n}{n} \to \mu$. Therefore, we are left to show

$$\sum_{k=1}^n \frac{Y_k - EY_k}{n} \to 0 \quad a.s.$$

By Kronecker's lemma, we only N.T.S.

$$\sum_{k=1}^n \frac{Y_k - EY_k}{k} \text{ converges a.s.}$$

Taking $A = 1$ in Kolmogorov's Three-series Theorem, (i) and (ii) therein are automatica satisfied. For (iii), recall that we have verified in the first proof of SLLN that

$$\sum_{k=1}^{\infty}\mathrm{Var}(\frac{Y_k - EY_k}{k}) \leq \sum_{k=1}^n \frac{E(Y_k^2)}{k^2} < \infty.$$

Theorem 2.5.8 (Marcinkiewicz-Zygmund SLLN). Let $X_1, X_2, \dots$ be i.i.d. with $EX_i = 0, E|X_i|^p < \infty$ for some $1 < p < 2$. Then with $S_n = \sum_{i=1}^n X_i$, we have

$$\frac{S_n}{n^{1/p}} \to 0 \quad a.s..$$

Proof. By Kronecker's lemma, it suffices to prove

$$\sum_{n=1}^{\infty}\frac{X_n}{n^{1/p}} \quad \text{converges a.s.}$$

We choose $A = 1$ in Komogorov's Three-series Theorem, define

$$Y_i = \frac{X_i}{i^{1/p}}1_{\{|X_i| \leq i^{1/p}\}},$$

and verify the three conditions below.
(i) We have

$$\sum_{n=1}^{\infty} P(|\frac{X_n}{n^{1/p}}| > 1) = \sum_{n=1}^{\infty} P(|X_1|^p > n) \leq E|X_1|^p < \infty.$$

(ii) We have

$$\sum_{i=1}^{\infty}|EY_i| = \sum_{i=1}^{\infty}\frac{1}{i^{1/p}}|EX_i 1_{\{|X_i| \leq i^{1/p}\}}|$$
$$= \sum_{i=1}^{\infty}\frac{1}{i^{1/p}}|EX_i 1_{\{|X_i| > i^{1/p}\}}| \quad \text{(from } EX_i = 0)$$
$$\leq \sum_{i=1}^{\infty}\sum_{j=i}^{\infty}\frac{1}{i^{1/p}}E[|X_1|1_{\{j \leq |X_1|^p < j+1\}}] \quad \text{(from the i.i.d. assumption)}$$
$$= \sum_{j=1}^{\infty}\sum_{i=1}^{j}\frac{1}{i^{1/p}}E[|X_1|1_{\{j \leq |X_1|^p < j+1\}}] \quad \text{(Fubini)}$$
$$\leq C_p\sum_{j=1}^{\infty}j^{1-\frac{1}{p}}E[|X_1|1_{\{j \leq |X_1|^p < j+1\}}] \quad \text{(sum of geometric series)}$$
$$\leq C_p\sum_{j=1}^{\infty}E\left\{(|X_1|^p)^{1-\frac{1}{p}}[|X_1|1_{\{j \leq |X_1|^p < j+1\}}]\right\}$$
$$\leq C_pE[|X_1|^p] < \infty.$$

(iii) We have, similarly as in verifying (ii),

$$\sum_{i=1}^{\infty}E[Y_i^2] = \sum_{i=1}^{\infty}\frac{1}{i^{2/p}}E(X_i^2 1_{\{|X_i| \leq i^{1/p}\}})$$
$$= \sum_{i=1}^{\infty}\sum_{j=1}^{i}\frac{1}{i^{2/p}}E[X_1^2 1_{\{j-1 < |X_1|^p \leq j\}}]$$
$$= \sum_{j=1}^{\infty}\sum_{i=j}^{\infty}\frac{1}{i^{2/p}}E[X_1^2 1_{\{j-1 < |X_1|^p \leq j\}}]$$
$$\leq C_p\sum_{j=1}^{\infty}j^{1-\frac{2}{p}}E[X_1^2 1_{\{j-1 < |X_1|^p \leq j\}}]$$
$$\leq C_p\sum_{j=1}^{\infty}E\left\{((|X_1|+1)^p)^{1-\frac{2}{p}}[X_1^2 1_{\{j \leq |X_1|^p < j+1\}}]\right\}$$
$$\leq C_pE[(|X_1|+1)^p] < \infty.$$

$\mu(\{a\}) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}e^{-ita}\varphi(t)dt$
$I_T = \frac{1}{2T}\int_{-T}^{T}e^{-ita}\varphi(t)dt$
$= \frac{1}{2T}\int_{-T}^{T}\int_{-\infty}^{+\infty}e^{it(x-a)}\mu(dx)dt$
Fubini $= \frac{1}{2T}\int_{-\infty}^{+\infty}\int_{-T}^{T}\cos(t(x-a)) + i\sin(t(x-a))dt\,\mu(dx)$
$= \frac{1}{2T}\int_{-\infty}^{+\infty}\int_{-T}^{T}\cos(\ )dt\,\mu(dx)$
$|\int_0^x \cos y\,dy| = |\sin y| \leq |x|$, $|\frac{1}{2T}\int_{-T}^{T}\cos(t(x-a))dt| \leq 1$
By DCT $\lim_{T\to\infty} I_T = \int_{-\infty}^{+\infty}\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\ \ \mu(dx)$
① $x \neq a$ $= \lim_{T\to\infty}\frac{\sin((x-a)T)}{(x-a)T} = 0$
② $x = a$ $= 1$
By ① and ②, $\lim_{T\to\infty} I_T = \mu(\{a\})$
THM3.3.5 If $\int_{-\infty}^{+\infty}|\varphi_X(t)|dt < \infty$
X conti. bounded conti. pdf
$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-itx}\varphi_X(t)dt$
$|\frac{e^{-ita} - e^{-itb}}{it}| = |\int_a^b e^{-ity}dy| \leq |b-a|$
$\mu(a,b) + \frac{1}{2}\mu(\{a,b\}) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt$
$\leq \frac{b-a}{2\pi}\int_{-\infty}^{+\infty}|\varphi(t)|dt$
$\mu(x, x+h) = \frac{1}{2\pi}\int\frac{e^{-itx} - e^{-it(x+h)}}{it}\varphi(t)dt$
$= \frac{1}{2\pi}\int(\int_x^{x+h}e^{-ity}dy)\varphi(t)dt$
$= \int_x^{x+h}\frac{1}{2\pi}\int e^{-ity}\varphi(t)dt\,dy$
$f(y) = \frac{1}{2\pi}\int e^{-ity}\varphi(t)dt$
By DCT f continuous

Central Limit Theorem

Sec1. Convergence in Distribution

Thm . (Stirling's Formula) $n! \sim n^n e^{-n}\sqrt{2\pi n}$ as $n \to \infty$.

Def. d.f.s $F_n \Rightarrow F$ weakly conv, if $F_n(y) \to F(y)$ at $\forall$ cont-point y of F

Fact. i) if $X_n \to X$ in p., then $X_n \Rightarrow X$. ii) if $X_n \Rightarrow c$, then $X_n \to c$ in p

Thm 3.2.2. (Skorokhod's Theorem) if $F_n \Rightarrow F_\infty$, then $\exists\ Y_n$ on the same prob-space, Y_n has d.f. F_n and $Y_n \to Y_\infty$ a.s.

Proof. Let $\Omega = (0,1)$, $\mathcal{F} = \{\text{Borel sets}\}$, $P =$Lebesgue measure. Define $Y_n : \Omega \to \mathbb{R}$ to be

$$Y_n(\omega) = F_n^{-1}(\omega),$$

where

$$F_n^{-1}(\omega) := \inf\{y : F_n(y) \geq \omega\} = \sup\{y : F_n(y) < \omega\}.$$

Recall from Chapter 1 that the d.f. of the above constructed Y_n is F_n. Let Ω_0 consists of those $x \in (0,1)$ such that the preimage of F at x is either empty or is a unique real number. Then Ω_0^c is countable; hence Ω_0 has probability 1. Moreover, it follows from simple calculus that $Y_n(x) \to Y_\infty(x)$ for all $x \in \Omega_0$. □

Slutsky's Theorem. If $X_n \Rightarrow X$ and $Y_n \to c$ in probability, where c is a constant. Suppose the random variables are defined on the same probability space. Then

$$X_n + Y_n \Rightarrow X + c, \quad X_nY_n \Rightarrow cX.$$

Proof. Let $x - c$ be a continuity point of F_X (so that x is a continuity point of F_{X+c}). Choose a decreasing sequence of $\epsilon \downarrow 0$ such that $x - c + \epsilon$ is a continuity point of F_x for every ϵ in the sequence (we can do this because continuity points are dense on $\mathbb{R}$). We have

$$P(X_n + Y_n \leq x) \leq P(|Y_n - c| \geq \epsilon) + P(X_n + Y_n \leq x, |Y_n - c| < \epsilon)$$
$$\leq P(|Y_n - c| \geq \epsilon) + P(X_n \leq x - c + \epsilon)$$
$$\to P(X + c \leq x + \epsilon) \quad \text{as } n \to \infty,$$
$$P(X_n + Y_n \leq x) \leq P(X + c \leq x).$$

Following similar arguments, we can prove the lower bound

$$P(X_n + Y_n \leq x) \geq P(X + c \leq x).$$

This shows $X_n + Y_n \Rightarrow X + c$.
To prove $X_nY_n \Rightarrow cX$ for the case $c > 0$, we use instead, for $0 < \epsilon < c$,

$$P(X_nY_n \leq x) \leq P(|Y_n - c| > \epsilon) + P(X_n \leq \frac{x}{c - \epsilon})$$

Theorem 3.2.3. $X_n \Rightarrow X$ if and only if $\forall\ g : \mathbb{R} \to \mathbb{R}$ bounded and continuous, we have $Eg(X_n) \to Eg(X)$.

Proof. "$\Longleftarrow$": Let F_n (F resp.) be the distribution function of X_n (X resp.). For any continuity point x of F and $\epsilon > 0$, define a continuous function

$$g(y) = \begin{cases} 1 & \text{if } y \leq x \\ 0 & \text{if } y \geq x + \epsilon \\ \text{linear} & \text{if } x \leq y \leq x + \epsilon. \end{cases}$$

We have

$$F_n(x) = E1_{\{X_n \leq x\}} \leq Eg(X_n)$$
$$\xrightarrow{\text{condition}} Eg(X) \leq E1_{\{X \leq x+\epsilon\}} = F(x + \epsilon) \downarrow F(x), \text{ as } \epsilon \downarrow 0.$$

Therefore, $\limsup_{n\to\infty}F_n(x) = F(x)$. Lower bound is proved similarly.
"$\Longrightarrow$": By Skorokhod's theorem, we can construct $Y_n, n \geq 1$ and Y on the same probability space such that Y_n has d.f. F_n for all n, Y has d.f. F and $Y_n \to Y$ a.s. We have

$$Eg(X_n) = Eg(Y_n) \xrightarrow{BCT} Eg(Y) = Eg(X).$$

Theorem (CLT assuming finite third moment). Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables such that $EX_i = \mu, Var(X_i) = \sigma^2$ and $E|X_i|^3 < \infty$. Let

$$W_n = \sum_{i=1}^n \frac{X_i - \mu}{\sigma\sqrt{n}}.$$

Then

$$W_n \xrightarrow{d} Z \sim N(0,1).$$

Proof. Let $g : \mathbb{R} \to \mathbb{R}$ be any bounded continuous function with bounded and continuous derivatives up to the third order. By the above remark, it suffices to prove $Eg(W_n) \to Eg(Z)$. For each $i = 1, \dots, n$, let

$$\xi_i = \frac{X_i - \mu}{\sigma\sqrt{n}}.$$

Define $\eta_1, \dots, \eta_n$ on the same probability space such that $\{\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_n\}$ are independent and $\eta_i \sim N(0, \frac{1}{n})$ for all i. Note that

$$E\xi_i = E\eta_i = 0, \ E\xi_i^2 = E\eta_i^2 = \frac{1}{n}, \ E|\xi_i|^3, E|\eta_i|^3 \leq \frac{C}{n^{3/2}}. \quad (1.1)$$

Note also that $\sum_{i=1}^n \eta_i \sim N(0,1)$. By using the telescoping sums of Taylor's expansion, we have

$$Eg(W_n) - Eg(Z)$$
$$= \sum_{k=1}^n\{Eg(\xi_1 + \dots + \xi_{k-1} + \xi_k + \eta_{k+1} + \dots + \eta_n) - Eg(\xi_1 + \dots + \xi_{k-1} + \eta_k + \eta_{k+1} + \dots + \eta_n)\}$$
$$= E\sum_{k=1}^n\left\{\left[g(V_k) + \xi_kg'(V_k) + \frac{\xi_k^2}{2}g''(V_k) + O(|\xi_k|^3)\right] - \left[g(V_k) + \eta_kg'(V_k) + \frac{\eta_k^2}{2}g''(V_k) + O(|\eta_k|^3)\right]\right\}$$

where $V_k := \xi_1 + \dots + \xi_{k-1} + \eta_{k+1} + \dots + \eta_n$. Using independence and (1.1) for cancellation, we obtain

$$|Eg(W_n) - Eg(Z)| \leq Cn\frac{1}{n^{3/2}} \to 0.$$

Definition. The **characteristic function (ch.f.)** of a random variable X is defined to be

$$\varphi_X(t) := Ee^{itX} = E\cos(tX) + i \cdot E\sin(tX).$$

Properties. 1. $\varphi_X(0) = 1$, $|\varphi_X(t)| \leq 1$.
2. $\varphi_X(-t) = \overline{\varphi_X(t)}$. (conjugate)
3. $|\varphi_X(t+h) - \varphi_X(t)| \leq E|e^{ihX} - 1| \to 0$, as $h \to 0$. (by DCT). That is, $\varphi_X(t)$ is uniformly continuous.
4. $\varphi_{aX+b}(t) = e^{itb}\varphi_X(at)$.
5. If X_1 is independent of X_2, then $\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t)$.

Theorem 3.3.8. If $E(X^2) < \infty$, then

$$\varphi_X(t) = 1 + i \cdot tE(X) - \frac{t^2}{2}E(X^2) + o(t^2), \text{ as } t \to 0.$$

Proof. By Taylor's expansion,

$$\varphi_X(t) = Ee^{itX} = 1 + EiXe^{itX}|_{t=0} \cdot t + E\frac{(iX)^2}{2}e^{itX}|_{t=0} \cdot t^2 + \text{error},$$

Theorem 3.3.4 (The Inversion Formula). Let $\varphi_X(t) = E(e^{itX})$ be the ch.f. of a random variable X. Then for any $a < b$, we have

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\varphi_X(t)dt = P(a < X < b) + \frac{1}{2}P(X = a) + \frac{1}{2}P(X = b).$$

The existence of the limit is part of the statement of the theorem.

Proof. First note that

$$|\frac{e^{-ita} - e^{-itb}}{it}| = |\int_a^b e^{-ity}dy| \leq b - a.$$

$$\lim_{T\to\infty}\int_0^T\frac{\sin(tc)}{t}dt = \begin{cases} \frac{\pi}{2} & c > 0, \\ 0 & c = 0, \\ -\frac{\pi}{2} & c < 0, \end{cases} \qquad |\int_0^T\frac{\sin(tc)}{t}dt| \leq 4.$$

By Fubini's theorem, for any T,

$$\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\varphi_X(t)dt$$
$$= E\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{it(X-a)} - e^{it(X-b)}}{it}dt$$
$$= E\frac{1}{2\pi}\int_{-T}^{T}\left[\frac{\cos(it(X-a)) - \cos(it(X-b))}{it} + \frac{\sin(t(X-a)) - \sin(t(X-b))}{t}\right]dt \quad (2.1)$$
$$= E\frac{1}{\pi}\int_0^T\left[\frac{\sin(t(X-a)) - \sin(t(X-b))}{t}\right]dt,$$

Theorem 3.3.6. (i) If $X_n \xrightarrow{d} X$, then $\varphi_{X_n}(t) \to \varphi_X(t)$, $\forall\ t \in \mathbb{R}$.
(ii) If $\varphi_{X_n}(t) \to \varphi(t)$, $\forall\ t \in \mathbb{R}$ and φ is continuous at 0, then φ is a ch.f. of some random variable X and $X_n \xrightarrow{d} X$.

Proof sketch. (i) follows from the Skorokhod's theorem and BCT.
For (ii), the continuity of φ at 0 implies tightness of the sequence X_n. Therefore, every subsequential vague limit is a distribution function. By the condition $\varphi_{X_n}(t) \to \varphi(t)$, every subsequential weak limit is the same distribution function having the ch.f. $\varphi(t)$. This implies the whole sequence converges weakly to the limit. □

Theorem. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables such that $EX_i = \mu, Var(X_i) = \sigma^2$. Let

$$W_n = \sum_{i=1}^n \frac{X_i - \mu}{\sigma\sqrt{n}}.$$

Then

$$W_n \xrightarrow{d} Z \sim N(0,1).$$

Proof. N.T.S. $Ee^{itW_n} \to e^{-\frac{t^2}{2}}$ for all $t \in \mathbb{R}$. We have, by the expression of W_n and independence,

$$Ee^{itW_n} = E\exp(it(\frac{X_1 - \mu}{\sigma\sqrt{n}} + \dots + \frac{X_n - \mu}{\sigma\sqrt{n}})) = \prod_{j=1}^n Ee^{i(\frac{t}{\sqrt{n}})(X_j - \mu)}$$
$$= \prod_{j=1}^n\left[1 + i\frac{t}{\sigma\sqrt{n}}E(X_j - \mu) - \frac{t^2}{2\sigma^2 n}E(X_j - \mu)^2 + o(\frac{t^2}{n})\right] \quad \text{(from Theorem 3.3.8)}$$
$$= \left[1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right]^n$$
$$\to e^{-t^2/2}.$$

Random Walks

Sec1. Random Walks

Def. if r-v-s X_i i.i.d. $S_n=\sum_i X_i$, then $\{S_n\}$ is random walk with $S_0=0$.

Thm 4.1.1. (Hewitt-Savage 0-1 Law)
if A is permutable (Not change under finite perm-), then $P(A)=0$ or 1.

Thm 4.1.2. For random walk on $\mathbb{R}$, one of follow- has prob- 1:
i) $S_n=0$ ii) $S_n\to\infty$ iii) $S_n\to-\infty$ iv) $-\infty=\liminf S_n<\limsup S_n=\infty$.

Proof. By the 0-1 law, $\{\limsup_{n\to\infty}S_n\ge c\}$ has probability 0 or 1. This mean with probability 1, $\limsup_{n\to\infty}S_n$ equals some value in $\mathbb{R}\cup\{-\infty,\infty\}$. Similarly, $\liminf_{n\to\infty}S_n$ equals some value in $\mathbb{R}\cup\{-\infty,\infty\}$. If $\limsup_{n\to\infty}S_n=c$ for some $c\in\mathbb{R}$, then from

$$\limsup_{n\to\infty}S_n\overset{d}{=}\limsup_{n\to\infty}(S_{n+1}-X_1)=\limsup_{n\to\infty}S_{n+1}-X_1=\limsup_{n\to\infty}S_n-X_1,$$

we must have $X_1=0$ and it belongs to case (i). Other cases correspond to other combinations of possibilities of $\limsup_{n\to\infty}S_n$ and $\liminf_{n\to\infty}S_n$. □

Def. $\mathcal{F}_0=\{\varnothing,\Omega\},\mathcal{F}_n=\sigma(X_1,\ldots,X_n)$ seq- of incr- σ-fields is filtration.

Def. rand-time $\tau\in\mathbb{R}^+\cup\{\infty\}$ is stopping time wrt $\{\mathcal{F}_n\}$ if $\{\tau=n\}\in\mathcal{F}_n$.

Fact. i) $\{\tau=n\}\in\mathcal{F}_n$ ii) $\{\tau\le n\}\in\mathcal{F}_n$ iii) $\{\tau\ge n+1\}\in\mathcal{F}_n$ are equival-.

Fact. i) $\tau_1\vee\tau_2$ ii) $\tau_1\wedge\tau_2$ iii) $\tau_1+\tau_2$ iv) $\tau_1\times\tau_2$ are stopping times.

Thm 4.1.5. (Wald's eq) i.i.d. if $\mathbb{E}[X_1],\mathbb{E}[\tau]\le\infty$, then $\mathbb{E}[S_\tau]=\mathbb{E}[X_1]\mathbb{E}[\tau]$.

Thm 4.1.6. (W's 2eq) i.i.d.$\sim(0,\sigma^2)$, if $\mathbb{E}[\tau]<\infty$, then $\mathbb{E}[S_\tau^2]=\sigma^2\mathbb{E}[\tau]$.

$$\begin{aligned}E(S_\tau)=&E(\sum_{i=1}^{\tau}X_i)\\=&E(\sum_{i=1}^{\infty}X_i1_{\{\tau\ge i\}})\\=&\sum_{i=1}^{\infty}E(X_i1_{\{\tau\ge i\}})\\=&\sum_{i=1}^{\infty}E(X_i)E(1_{\{\tau\ge i\}})\\=&E(X_1)E(\tau).\end{aligned}$$

$$S^2_{\tau\wedge n}=S^2_{\tau\wedge(n-1)}+(2X_nS_{n-1}+X_n^2)1_{\{\tau\ge n\}}.$$

$$\begin{aligned}ES^2_{\tau\wedge n}=&ES^2_{\tau\wedge(n-1)}+E\left(2X_nS_{n-1}1_{\{\tau\ge n\}}\right)+E[X_n^21_{\{\tau\ge n\}}]\\=&ES^2_{\tau\wedge(n-1)}+\sigma^2P(\tau\ge n)\\=&\cdots\\=&\sigma^2\sum_{i=1}^{n}P(\tau\ge i).\end{aligned}$$

Letting $n\to\infty$ gives the result.

Example. Let $X_1,X_2,\ldots$ be i.i.d. $\sim$ Uniform(0,1). $S_n=X_1+\cdots+X_n$. Let $\tau:=\inf\{n:S_n>1\}$. Then

$$E(\tau)=e,\quad E(S_\tau)=\frac{e}{2}.$$

Proof. We have

$$\begin{aligned}P(\tau>n)=&P(S_n\le1)\\=&\int_0^1\cdots\int_0^1 1_{\{x_1+\cdots+x_n\le1\}}dx_1\cdots dx_n &&\text{(from the uniform distribution)}\\=&\int_0^1\int_0^{y_n}\cdots\int_0^{y_2}dy_1dy_2\cdots dy_n &&\text{(by a change of variable)}\\=&\frac{1}{n!}.\end{aligned}$$

This implies

$$E(\tau)=\sum_{n=0}^{\infty}P(\tau>n)=e.$$

By Wald's first equation,

$$E(S_\tau)=E(X_1)E(\tau)=\frac{e}{2}.$$

Example. Let $X_1,X_2,\ldots$ be i.i.d. with $P(X_1=1)=P(X_1=-1)=\frac{1}{2}$. Let $S_n=X_1+\cdots+X_n$ (SRW). Let a,b be two integers with $a<0<b$. Let

$$N:=\inf\{n:S_n\notin(a,b)\}.$$

Then

1. $E(N)<\infty$,
2. $S_N=a$ or b,
3. $P(S_N=a)=\frac{b}{b-a},\quad P(S_N=b)=\frac{-a}{b-a}$,
4. $E(N)=E(S_N^2)=(-a)b$.

Proof. For any positive integer k, by dividing the interval $(0,k(b-a))$ into k subintervals of equal length and considering an extreme case behavior (keep going upwards) of the random walk within each subinterval, we obtain

$$P(N>k(b-a))\le(1-\frac{1}{2^{b-a}})^k.$$

This implies 1.
2. is obvious.
3. follows from Wald's first equation.
4. follows from Wald's second equation. □

Sec2. Recurrence v.s. Transience

$$\tau_1=\inf\{m\ge1:S_m=0\},$$
$$\tau_n=\inf\{m>\tau_{n-1}:S_m=0\}.$$

Theorem. The following are equivalent:
(i) $P(\tau_1<\infty)=1$.
(ii) $P(\tau_n<\infty)=1,\ \forall\ n=1,2,3,\ldots$.
(iii) $P(S_m=0\ i.o.)=1$.
(iv) $\sum_{m=1}^{\infty}P(S_m=0)=\infty$.

Proof. We have

$$\begin{aligned}P(\tau_2<\infty)=&P(\tau_1<\infty,\tau_2-\tau_1<\infty)\\=&\sum_{m,n=1}^{\infty}P(\tau_1=m,\tau_2-\tau_1=n)\\=&\sum_{m,n=1}^{\infty}P(X_1+\cdots+X_m=0,X_1+\cdots+X_u\ne0,\forall\,1\le u<m;\\&\qquad X_{m+1}+\cdots+X_{m+n}=0,X_{m+1}+\cdots+X_{m+v}\ne0,\forall\,1\le v<n)\\=&\sum_{m,n=1}^{\infty}P(\tau_1=m)P(\tau_1=n) &&\text{(by i.i.d. assumption)}\\=&(P(\tau_1<\infty))^2.\end{aligned}$$

Similarly, we can prove

$$P(\tau_n<\infty)=(P(\tau_1<\infty))^n. \tag{2.1}$$

$$\sum_{m=0}^{\infty}P(S_m=0)=\sum_{m=0}^{\infty}E1_{\{S_m=0\}}=E\sum_{m=0}^{\infty}1_{\{S_m=0\}}=\sum_{n=0}^{\infty}(P(\tau_1<\infty))^n.$$

Theorem 4.2.3. SRW is recurrent in $\mathbb{R}^1$ and $\mathbb{R}^2$ and is transient in $\mathbb{R}^d,d\ge3$.

Proof. In $\mathbb{R}^1$,

$$\begin{aligned}\sum_{m=1}^{\infty}P(S_m=0)=&\sum_{n=1}^{\infty}P(S_{2n}=0) &&\text{(can only return to 0 a}\\=&\sum_{n=1}^{\infty}\binom{2n}{n}(\frac{1}{2})^{2n} &&\text{(co}\\\sim&\sum_{n=1}^{\infty}\frac{\sqrt{2\pi2n}(\frac{2n}{e})^{2n}}{(\sqrt{2\pi n}(\frac{n}{e})^n)^2}\frac{1}{2^{2n}} &&\text{(Stirli}\\=&\sum_{n=1}^{\infty}\frac{1}{\sqrt{\pi n}}=\infty;\end{aligned}$$

hence recurrent by the previous theorem. By $\mathbb{R}^2$, similar calculation yields

$$P(S_{2n}=0)\asymp\frac{1}{n},$$

still sum to infinity; hence still recurrent.

In $\mathbb{R}^3$, more complicated combinatorics give $P(S_{2n}=0)\asymp\frac{1}{n^{3/2}}$ summing t

Theorem (Reflection Principle). Let $X_1,X_2,\ldots$ be i.i.d. with $P(X_1=1)=P(X_1=-1)=\frac{1}{2}$, $S_n=X_1+\cdots+X_n$. For any positive integer b, we have

$$P(\max_{1\le k\le n}S_k\ge b)=2P(S_n>b)+P(S_n=b).$$

Proof. We have

$$\begin{aligned}&P(\max_{1\le k\le n}S_k\ge b)\\=&P(\max_{1\le k\le n}S_k\ge b,S_n>b)+P(\max_{1\le k\le n}S_k\ge b,S_n<b)+P(\max_{1\le k\le n}S_k\ge b,S_n=b).\end{aligned}$$

The first two terms are equal by reflecting the random walk trajectory along the horizontal line $y=b$ after it first hits level b. They both equal $P(S_n>b)$. The third term is simply $P(S_n=b)$. This gives the result. □

Martingale

Sec1. Conditional Expectation

Def. $(\Omega,\mathcal{F},\mathcal{P})$, $\mathbb{E}[|X|]<\infty$, σ-field $\mathcal{A}\subset\mathcal{F}$, $\mathbb{E}[X|\mathcal{A}]$ cond-expectation, if i) $\mathbb{E}[X|\mathcal{A}]$ is $\mathcal{A}$-measurable, ii) $\forall\,A\in\mathcal{A}$, $\mathbb{E}[X1_A]=\mathbb{E}[\mathbb{E}[X|\mathcal{A}]1_A]$.

Prop. (a) $\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X|\mathcal{A}]]$.
(b) $|\mathbb{E}[X|\mathcal{A}]|\le\mathbb{E}[|X||\mathcal{A}]$.
(c) if $X\in\mathcal{A}$, then $\mathbb{E}[X|\mathcal{A}]=X$.
(d) if X is independent of $\mathcal{A}$, then $\mathbb{E}[X|\mathcal{A}]=\mathbb{E}[X]$.
(e) (linearity) if $\mathbb{E}[X],\mathbb{E}[Y]<\infty$, then $\mathbb{E}[aX+Y|\mathcal{A}]=a\mathbb{E}[X|\mathcal{A}]+\mathbb{E}[Y|\mathcal{A}]$.
(f) (monotonicity) if $X\le Y$, then $\mathbb{E}[X|\mathcal{A}]\le\mathbb{E}[Y|\mathcal{A}]$.
Hint: consider $A=\{\mathbb{E}[X|\mathcal{A}]-\mathbb{E}[Y|\mathcal{A}]\ge\epsilon>0\}$.
(g) (cmct) if $X_n\ge0,X_n\uparrow X,\mathbb{E}[X]<\infty$, then $\mathbb{E}[X_n|\mathcal{A}]\uparrow\mathbb{E}[X|\mathcal{A}]$.
(h) (cFatou) if $X_n\ge0$, $\mathbb{E}[\underline{\lim}X_n]<\infty$, then $\underline{\lim}\mathbb{E}[X_n|\mathcal{A}]\ge\mathbb{E}[\underline{\lim}X_n|\mathcal{A}]$.
(i) (cdct) if $|X_n|\le Y,\mathbb{E}[Y]<\infty$, $X_n\to X$a.s, then $\mathbb{E}[X_n|\mathcal{A}]\to\mathbb{E}[X|\mathcal{A}]$ a.s.
(j) (cJen) if conv-φ, $\mathbb{E}[|X|],\mathbb{E}[|\varphi(X)|]<\infty$, then $\varphi(\mathbb{E}[X|\mathcal{A}])\le\mathbb{E}[\varphi(X)|\mathcal{A}]$.
(k) (cHol) if $p,q\ge1$ and $\frac{1}{p}+\frac{1}{q}=1$, $\mathbb{E}[|X|^p],\mathbb{E}[|Y|^q]<\infty$,
$\mathbb{E}[|XY||\mathcal{A}]\le(\mathbb{E}[|X|^p|\mathcal{A}])^{1/p}(\mathbb{E}[|Y|^q|\mathcal{A}])^{1/q}$.
Hint: suppose $\mathbb{E}[|X|^p|\mathcal{A}]\ge\epsilon>0$, take $|X'|=(|X|^p+\epsilon^p)^{1/p}$, DCT.
(l) (cMin) if $p\ge1,\mathbb{E}[|X|^p],\mathbb{E}[|Y|^p]<\infty$,
then $(\mathbb{E}[|X+Y|^p|\mathcal{A}])^{1/p}\le(\mathbb{E}[|X|^p|\mathcal{A}])^{1/p}+(\mathbb{E}[|Y|^p|\mathcal{A}])^{1/p}$
(m) (cMar) if $X\ge0,a>0$, then $P(X\ge a|\mathcal{A})\le\mathbb{E}[X|\mathcal{A}]/a$.
(n) if $\mathbb{E}[|X|],\mathbb{E}[|XY|]<\infty$ and $X\in\mathcal{A}$, then $\mathbb{E}[XY|\mathcal{A}]=X\mathbb{E}[Y|\mathcal{A}]$.

Proof. We proceed to verify that $XE(Y|\mathcal{A})$ satisfies (i) and (ii) in the definition of conditional expectation. (i) follows from the assumption that $X\in\mathcal{A}$. To verify (ii), we N.T.S. for any $A\in\mathcal{A}$,

$$E(XY1_A)=E[XE(Y|\mathcal{A})1_A]. \tag{1.1}$$

To verify (1.1), we proceed by considering X being an 1. indicator variable, 2. simple variable, 3. positive variable and 4. general variable as before.

1. Suppose $X=1_B$, $B\in\mathcal{A}$. Then

$$\begin{aligned}RHS(1.1)=&E[1_B1_AE(Y|\mathcal{A})]=E[E(Y|\mathcal{A})1_{A\cap B}]\\\overset{(ii)}{=}&E(Y1_{A\cap B})=E(1_BY1_A)=E(XY1_A)=LHS(1.1).\end{aligned}$$

2. Suppose $X=\sum_{i=1}^n b_i1_{B_i}$. Then

$$E(\sum_{i=1}^{n}b_i1_{B_i}Y|\mathcal{A})\overset{linearity}{=}\sum_{i=1}^{n}b_iE(1_{B_i}Y|\mathcal{A})\overset{1.}{=}\sum_{i=1}^{n}b_i1_{B_i}E(Y|\mathcal{A})=XE(Y|\mathcal{A}).$$

3. Use monotone convergence theorem of the conditional expectation above and approximate X by simple variables from below.
4. Write general X as X^+-X^-. □

(o) (Tower Property) if $\mathbb{E}[|X|]<\infty,\mathcal{A}_1\subset\mathcal{A}_2$, then
i) $\mathbb{E}[\mathbb{E}[X|\mathcal{A}_1]|\mathcal{A}_2]=\mathbb{E}[X|\mathcal{A}_1]$ ii) $\mathbb{E}[\mathbb{E}[X|\mathcal{A}_2]|\mathcal{A}_1]=\mathbb{E}[X|\mathcal{A}_1]$
$\forall\,A\in\mathcal{A}_1:\quad E(E(X|\mathcal{A}_1)1_A)=E(X1_A)=E(E(X|\mathcal{A}_2)1_A)$.
(p) (triangular eq) if $\mathbb{E}[X^2]<\infty$, then for $\forall\,Y\in\mathcal{A}$ with $\mathbb{E}[Y^2]<\infty$,
$\mathbb{E}[(X-\mathbb{E}[X|\mathcal{A}])^2]\le\mathbb{E}[(X-Y)^2]$.
(q) Var$(X)\ge\mathbb{E}[$Var$(X|\mathcal{A})]$.
(r) if $Z\perp\!\!\!\perp(X,Y)$, then $\mathbb{E}[X|Y,Z]=\mathbb{E}[X|Y]$.
Hint: $\mathcal{P}=\{B\cap C\}$, $\mathcal{L}=\{A\in\sigma(Y,Z):\mathbb{E}[X1_A]=\mathbb{E}[\mathbb{E}[X|Y]1_A]\}$.

Sec2. Martingale

Def. $\{S_n\}$ is martingale w.r.t. filtration $\{\mathcal{F}_n\}$ if
i) $\mathbb{E}[|S_n|]<\infty$, ii) $S_n\in\mathcal{F}_n$, iii) $\mathbb{E}[S_{n+1}|\mathcal{F}_n]=S_n$.

Fact. i) $\mathbb{E}[S_{n+m}|\mathcal{F}_n]=S_n$, ii) $\mathbb{E}[S_1]=\mathbb{E}[S_2]=$, iii) $\mathbb{E}[|S_1|]\le\mathbb{E}[|S_2|]\le$.

Def. martingale difference: $X_n=S_n-S_{n-1}$, and $S_0=0$.

Fact. i) $\mathbb{E}[X_i]=0$ for $i\ge2$. ii) $\mathbb{E}[S_n^2]=\mathbb{E}[X_1^2]+\cdots+\mathbb{E}[X_n^2]$.

Def. super-(sub-)martingale: iii) $\mathbb{E}[S_n|\mathcal{F}_{n-1}]\le(\ge)S_{n-1}$.

Remark: martingale is both super-/sub-martingale. $E(S_{n-1}X_n)=E(E(S_{n-1}X_n|\mathcal{F}_{n-1}))=0.$

Thm 5.2.3-4. (martingale transformation)
i) if $\{S_n,\mathcal{F}_n\}$ mt, φ convex, $\mathbb{E}[|\varphi(S_n)|]<\infty$, then $\{\varphi(S_n)\}$ is sub-mt.
ii) if $\{S_n\}$ sub-mt, φ convex $\uparrow$, $\mathbb{E}[|\varphi(S_n)|]<\infty$, then $\{\varphi(S_n)\}$ is sub-mt.

For case (1), we use Jensen's inequality to obtain

$$E(\varphi(S_n)|\mathcal{F}_{n-1})\ge\varphi(E(S_n|\mathcal{F}_{n-1}))=\varphi(S_{n-1}),$$

verifying (iii).

For case (2), we use Jensen's inequality and the additional assumption that φ is increasing to obtain

$$E(\varphi(S_n)|\mathcal{F}_{n-1})\ge\varphi(E(S_n|\mathcal{F}_{n-1}))\ge\varphi(S_{n-1}),$$

Theorem 5.2.8: Martingale Convergence Theorem. Suppose $(S_n,\mathcal{F}_n)$ is a submartingale and $\liminf E(S_n^+)<\infty$. Then S_n converges almost surely (a.s.) to a limit S with $E|S|<\infty$.

As a consequence, we recover the following result we learnt before.

Corollary 1. Suppose $X_1,X_2,\ldots$ are independent and $E(X_i)=0$. If $\sum_{i=1}^\infty E(X_i^2)<\infty$, then $S_n=\sum_{i=1}^n X_i$ converges a.s.

Proof of Corollary 1. Recall that mean zero independent sums S_n is a martingale (hence a submartingale). It suffices to verify the condition of Theorem 5.2.8. This follows from

$$E(S_n^+)\le E|S_n|\le\sqrt{E(S_n^2)}=\sqrt{\sum_{i=1}^{\infty}E(X_i^2)}<\infty.$$

Corollary 2. Suppose $X_1,X_2,\ldots$ are identically distributed (not necessarily independent) and $E|X_1|<\infty$. Let $\mathcal{F}_n=\sigma(X_1,\ldots,X_n)$. Let $S_1=X_1$ and $S_n=S_{n-1}+X_n-E(X_n|\mathcal{F}_{n-1}),n\ge2$. Then $(S_n,\mathcal{F}_n)$ is a martingale and

$$\frac{S_n}{n}\to0\quad\text{in probability.}$$

Proof of Corollary 2. We omit the straightforward verification that $(S_n,\mathcal{F}_n)$ is a martingale. Let $Y_i=X_iI(|X_i|\le i)$, $i\ge1$. Let

$$T_1=Y_1,\quad T_n=T_{n-1}+Y_n-E(Y_n|\mathcal{F}_{n-1}),n\ge2.$$

We have

$$\sum_{i=1}^{\infty}P(Y_i\ne X_i)=\sum_{i=1}^{\infty}P(|X_1|>i)<\infty\ \text{(from }E|X_1|<\infty) \tag{2.1}$$

and

$$\sum_{i=2}^{\infty}\frac{E[Y_i-E(Y_i|\mathcal{F}_{i-1})]^2}{i^2}\le\sum_{i=2}^{\infty}\frac{E(Y_i^2)}{i^2}\le CE|X_1|<\infty, \tag{2.2}$$

where we used the triangle inequality for conditional expectations in the first inequality and some result we proved before when we studied law of large numbers for the second inequality.

From the martingale convergence theorem and (2.2) above, we have

$$\sum_{i=2}^{n}\frac{Y_i-E(Y_i|\mathcal{F}_{i-1})}{i}\ \text{converges a.s.} \tag{2.3}$$

From (2.3) and the Kronecker's lemma, we have

$$\frac{1}{n}\sum_{i=2}^{n}(Y_i-E(Y_i|\mathcal{F}_{i-1}))\to0\ a.s. \tag{2.4}$$

(2.4) and (2.1) imply

$$\frac{1}{n}\sum_{i=2}^{n}(X_i-E(Y_i|\mathcal{F}_{i-1}))\to0\ a.s. \tag{2.5}$$

Corollary 2 follows from (2.5) and

$$\frac{1}{n}\sum_{i=2}^{n}E(Y_i|\mathcal{F}_{i-1})-\frac{1}{n}\sum_{i=2}^{n}E(X_i|\mathcal{F}_{i-1})\to0\ \text{in probability.} \tag{2.6}$$

$$\begin{aligned}E|E(X_i-Y_i|\mathcal{F}_{i-1})|\le&E|E(X_iI(|X_i|>i)|\mathcal{F}_{i-1})|\\\le&E|X_iI(|X_i|>i)|=E|X_1I(|X_1|>i)|\to0\end{aligned}$$

by the dominated convergence theorem.

Theorem 5.2.9. $(S_n,\mathcal{F}_n)$ is a supermartingale. If $S_n\ge0,\forall n\ge1$, then $S_n\to S$ a.s. and $ES<\infty$.

Proof. $-S_n\le0$ is a submartingale. $E((-S_n)^+)=0$. By the martingale convergence theorem, we have $S_n\to S$ a.s. Moreover, by Fatou and the definition of supermartingales,

$$ES=E(\lim S_n)\le\liminf ES_n\le ES_1.$$

Theorem 5.2.6. Let $\{\mathcal{F}_n,n\ge1\}$ be a filtration. Recall that $N\ge1$ is a **stopping time** i $\{N=n\}\in\mathcal{F}_n$, $\forall n\ge1$. If $(S_n,\mathcal{F}_n)$ is a submartingale, then

$$T_n:=S_{n\wedge N}$$

is a submartingale.

Proof. Let $X_1=S_1$ and $X_i=S_i-S_{i-1}$, $i\ge2$. Rewrite

$$T_n=S_{n\wedge N}=\sum_{i=1}^{n\wedge N}X_i=\sum_{i=1}^{n}X_iI(N\ge i).$$

(i) and (ii) in the definition of submartingales follow easily from the above expression. To prove (iii), we have

$$\begin{aligned}E(T_n|\mathcal{F}_{n-1})=&T_{n-1}+E(X_nI(N\ge n)|\mathcal{F}_{n-1})\\=&T_{n-1}+I(N\ge n)E(X_n|\mathcal{F}_{n-1})\ge T_{n-1}.\end{aligned}$$

$$\begin{aligned}ET_n=&\sum_{i=1}^{n}EX_iI(N\ge i)=EX_1+\sum_{i=2}^{n}E[I(N\ge i)E(X_i|\mathcal{F}_{i-1})]\\\le&EX_1+\sum_{i=2}^{n}E[E(X_i|\mathcal{F}_{i-1})]=EX_1+\sum_{i=2}^{n}E(X_i)=ES_n.\end{aligned}$$

Theorem 5.4.2 (Doob's inequality). $(S_n,\mathcal{F}_n)$ is a submartingale. Then for any $x>0$, we have

$$P(\max_{1\le k\le n}S_k\ge x)\le\frac{1}{x}E\Big[S_nI(\max_{1\le k\le n}S_k\ge x)\Big]\le\frac{E(S_n^+)}{x}.$$

Exercise: Show that Kolmogorov's maximal inequality follows from the above inequality.

Proof of Theorem 5.4.2. Let $N=\inf\{k:S_k\ge x\}$. N is a stopping time. Let

$$A=\{\max_{1\le k\le n}S_k\ge x\}=\{N\le n\}.$$

Then

$$1_A\le\frac{S_{n\wedge N}}{x}1_A.$$

We have, from (2.7),

$$\begin{aligned}P(A)=&E(1_A)\le\frac{1}{x}E(S_{n\wedge N}1_A)\\=&\frac{1}{x}\Big[E(S_{n\wedge N})-E(S_{n\wedge N})1_{A^c}\Big]\\\le&\frac{1}{x}\Big[E(S_n)-ES_n1_{A^c}\Big]\\=&\frac{1}{x}E(S_n1_A).\end{aligned}$$

Theorem 5.4.3 (L^p maximum inequality). For a submartingale S_n and any constant $p>1$, we have

$$E\Big[(\max_{1\le k\le n}S_k^+)^p\Big]\le\Big(\frac{p}{p-1}\Big)^pE\Big[(S_n^+)^p\Big].$$

Proof. Let $M>0$ be a constant. We have

$$\begin{aligned}&E\Big[(\max_{1\le k\le n}S_k^+\wedge M)^p\Big]\\=&\int_0^\infty p\lambda^{p-1}P(\max_{1\le k\le n}S_k^+\wedge M\ge\lambda)d\lambda\\\le&\int_0^\infty p\lambda^{p-1}\frac{1}{\lambda}E\Big[S_n^+I(\max_{1\le k\le n}S_k^+\wedge M\ge\lambda)\Big]d\lambda,\end{aligned} \tag{2.8}$$

where the equality is from the alternative expression of moments of positive random variables and the last inequality is trivial for $M<\lambda$ and from Doob's inequality applied to the submartingale S_n^+ for $M\ge\lambda$.

From (2.8) and direct integration, we have

$$\begin{aligned}&E\Big[(\max_{1\le k\le n}S_k^+\wedge M)^p\Big]\\\le&\frac{p}{p-1}ES_n^+(\max_{1\le k\le n}S_k^+\wedge M)^{p-1}\\\le&\frac{p}{p-1}(E(S_n^+)^p)^{1/p}(E(\max_{1\le k\le n}S_k^+\wedge M)^p)^{1/q},\end{aligned}$$

where $1/p+1/q=1$. The theorem is proved by moving the last factor on the right-hand side to the left. □

Theorem 5.4.5 (L^p convergence theorem). If (S_n) is a martingale and for some constant $p>1$, $\sup_kE|S_k|^p<\infty$, then S_n converges almost surely to a limit S and $E|S_n-S|^p\to0$.

Proof. The a.s. convergence follows from the martingale convergence theorem. We only need to show $E|S_n-S|^p\to0$.

Because (S_n) is a martingale, we have $(|S_n|)$ is a submartingale. From the previous theorem, we have

$$E(\max_{1\le k\le n}|S_k|)^p\le\Big(\frac{p}{p-1}\Big)^pE(|S_n|^p).$$

This means

$$E(\sup_{k\ge1}|S_k|)^p\le\Big(\frac{p}{p-1}\Big)^p\sup E(|S_n|^p),$$

which is finite by the assumption.

Because

$$|S_n-S|^p\le(2\sup_{k\ge1}|S_k|)^p,$$

we have, by the dominated convergence theorem,

$$E|S_n-S|^p\to0.$$

Theorem 3.4.5 (The Lindeberg-Feller Theorem). Assume for each n, $\xi_{n1},\xi_{n2},\ldots,\xi_{nn}$ are independent with $E\xi_{ni}=0$ for all i and $E\sum_{i=1}^n\xi_{ni}^2=1$. If

$$\forall\,\epsilon>0,\quad\sum_{i=1}^{n}E\xi_{ni}^21_{\{|\xi_{ni}|>\epsilon\}}\to0,\qquad\text{(Lindeberg's Condition)}$$

then

$$\sum_{i=1}^{n}\xi_{ni}\overset{d}{\to}N(0,1).$$

Proof of the Lindeberg-Feller theorem. Let

$$\varphi_n(t)=Ee^{it\sum_{i=1}^n\xi_{ni}}.$$

From the proof of Theorem 3.3.8, we have

$$\begin{aligned}\varphi_n(t)=&\prod_{i=1}^{n}Ee^{it\xi_{ni}}\\=&\prod_{i=1}^{n}E\Big[1+it\xi_{ni}-\frac{t^2}{2}\xi_{ni}^2+O(t^2\xi_{ni}^21_{\{|\xi_{ni}|>\epsilon\}})+O(t^3|\xi_{ni}|^31_{\{|\xi_{ni}|\le\epsilon\}})\Big]\\=&\prod_{i=1}^{n}\Big[1-\frac{t^2}{2}E\xi_{ni}^2+O(t^2E\xi_{ni}^21_{\{|\xi_{ni}|>\epsilon\}})+O(t^3\epsilon E|\xi_{ni}|^2)\Big]\\\to&\prod_{i=1}^{n}e^{-\frac{t^2}{2}E\xi_{ni}^2}=e^{-t^2/2},\end{aligned}$$

where we used Lemma 3.4.3: Let $z_1,\ldots,z_n$ and $w_1,\ldots,w_n$ be complex numbers with $|z_i|\le1,|w_i|\le1$ for all i. Then

$$\Big|\prod_{i=1}^{n}z_i-\prod_{i=1}^{n}w_i\Big|\le\sum_{i=1}^{n}|z_i-w_i|.$$

Question 3: Pólya's urn. A bag contains red and blue balls, with initially r red and b blue where $rb>0$. A ball is drawn from the bag at random, its colour noted, and then returned to the bag together with a new ball of the same colour. Let R_n be the number of red balls after n such operations.
(a) Show that $\{Y_n=R_n/(n+r+b),n\geqslant0\}$ is a martingale.
(b) Show that Y_n converges almost surely.
(c) Let T be the number of balls drawn until the first blue ball appears, and suppose that $r=b=1$. Compute $E[T/(T+2)]$.
(d) Suppose $r=b=1$. Show that $P(Y_n\geqslant\frac{3}{4}$ for some $n)\leqslant\frac{2}{3}$.

[Note: You may use the following version of Doob's Optional Stopping Theorem: If the sequence $\{Y_n,n\geqslant0\}$ is a bounded martingale and T is a stopping time, then the expected value of Y_T is Y_0.]

(a) $X_{n+1}=R_{n+1}-R_n=\begin{cases}0, & p=\frac{n+r+b-R_n}{n+r+b}\\1, & p=\frac{R_n}{n+r+b}\end{cases}$

$E(Y_{n+1}|\mathcal{F}_n)=E\left(\frac{R_n+X_{n+1}}{n+r+b+1}\Big|\mathcal{F}_n\right)=\frac{1}{n+r+b+1}\left(\frac{R_n(n+r+b-R_n)}{n+r+b}+\frac{(R_n+1)R_n}{n+r+b}\right)=\frac{R_n}{n+r+b}$

(b) $|Y_n|\in[0,1]\Rightarrow\liminf E(Y_n^+)<\infty\overset{\text{THM5.2.8}}{\Rightarrow}Y_n\overset{a.s.}{\to}Y$.
submartingale

(c) $T=\inf\{n:r+b+n-R_n=2\}$. Since $\{T=n\}\in\mathcal{F}_n$, T stopping time.
$E(Y_T)=E(\frac{R_T}{T+2})=E(\frac{T}{T+2})=Y_0=\frac{1}{2}$
(d) $P(\max_{1\le k\le n}Y_k\ge\frac{3}{4})\le\frac{E(Y_n)}{\frac{3}{4}}=\frac{4Y_0}{3}=\frac{2}{3}$

[Question] $P(X_k=\pm k)=\frac{1}{2k^2}$, $P(X_k=\pm1)=\frac{1}{2}(1-\frac{1}{k^2})$, then
$\begin{cases}S_n/(\sum_{i=1}^n X_i^2)^{\frac{1}{2}}\to N(0,1)\\S_n/\sqrt{n}\log n\overset{a.s.}{\to}0\end{cases}$

Proof: (a) $EX_i=0$, $EX_i^2=1-\frac{1}{2i^2}$. Let $\xi_{ni}=X_i/(\sum_{i=1}^n EX_i^2)^{\frac{1}{2}}$
Then $E\xi_{ni}=0$. $\sum_{i=1}^n E\xi_{ni}^2=1$.
For $\forall\varepsilon>0$, $\sum_{i=1}^n E\xi_{ni}^2I(|\xi_{ni}|>\varepsilon)=\frac{\sum_{i=1}^n EX_i^2I\{|X_i|>(\sum_{i=1}^n EX_i^2)^{\frac{1}{2}}\varepsilon\}}{\sum_{i=1}^n EX_i^2}\to0$
By Lindeberg-Feller CLT,
$\sum_{i=1}^n\xi_{ni}=\frac{\sum_{i=1}^n X_i}{(\sum_{i=1}^n EX_i^2)^{\frac{1}{2}}}\overset{d}{\to}N(0,1)$

Now we N.T.S. $(\sum_{i=1}^n X_i^2-E\sum_{i=1}^n X_i^2)/\sum_{i=1}^n EX_i^2\overset{P}{\to}0$.
By SLLN, $(\sum_{i=1}^n X_i^2-\sum_{i=1}^n EX_i^2)/n\to0$ a.s.
and $\sum_{i=1}^n EX_i^2\sim n$. Then $\sum_{i=1}^n X_i^2/\sum_{i=1}^n EX_i^2\overset{P}{\to}1$.
So $S_n/(\sum_{i=1}^n X_i^2)^{\frac{1}{2}}\to N(0,1)$

(b) To prove $S_n/\sqrt{n}\log n\to0$, it suffices to show $\sum_{n=1}^\infty(S_n/\sqrt{n}\log n)<\infty$
Use Three-series thm with $A=1$, $Y_i=X_iI(|X_i|>1)$
Then (i) $\sum_{n=1}^\infty P(|X_n|>1)<\infty$; (ii) $\sum_{n=1}^\infty EY_n=0$. $i\log i$
(iii) $\sum_{n=1}^\infty EY_n^2\le\sum_{n=1}^\infty\frac{1}{n(\log n)^2}<\infty$. Then $\frac{S_n}{n\log n}\to0$. a.s.

[Question] $X_i\overset{iid}{\sim}U[-1,1]$. (a) $W_n=\frac{1}{n}\sum_{1\le i\le j\le n}X_iX_j\overset{d}{\to}N(0,\frac{1}{9})$
(b) $\frac{S_n}{\sqrt{n}\log n}\overset{a.s.}{\to}0$ (c) $P(|S_n|>\sqrt{n}t)\le2\exp(-\frac{t^2}{2})$

Proof: (a) Let $\xi_{nij}=3X_iX_j/(n(n-1))^{\frac{1}{2}}$, $1\le i\le j\le n$.
Then $\sum_{1\le i\le j\le n}E\xi_{nij}=0$, $\sum_{i,j}E\xi_{nij}^2=1$.
$\sum_{i,j}E\xi_{nij}^2I(|\xi_{nij}|>\varepsilon)\to0$. By Linderberg-Feller CLT.
we have $\sum_{1\le i\le j\le n}\xi_{nij}\overset{d}{\to}N(0,1)$
$W_n=\frac{1}{n}\sum_{i,j}X_iX_j=\frac{(n(n-1))^{\frac{1}{2}}}{3n}\sum_{1\le i,j\le n}\xi_{nij}\overset{d}{\to}N(0,\frac{1}{9})$
(b) See (b)↑.
(c) Use Hoeffding's inequality, suppose $X_1,\ldots,X_n$ indep.
$P(X_i\in[a_i,b_i])=1$. then $P(|\bar{X}-E(\bar{X})|\ge t)\le$
$2\exp(-2t^2n^2/\sum_{i=1}^n(b_i-a_i)^2)$
Let $a_i=-1,b_i=1$, $E\bar{X}=0$, $t'=\frac{t}{\sqrt{n}}$.
$P(|S_n/n|>\frac{t}{\sqrt{n}})\le2\exp(-\frac{t^2}{2})$

[Question] S_n submartingale. (a) $ES_{n\wedge T}\le ES_n$.
$S_{n\wedge T}=\sum_{i=1}^n X_iI\{T\ge i\}$; $ES_{n\wedge T}=\sum_{i=1}^n E(E(X_iI\{T\ge i\}|\mathcal{F}_{i-1}))+EX_1$
$=\sum_{i=2}^n E(X_iP(T\ge i|\mathcal{F}_{i-1}))\le\sum_{i=1}^n X_i=S_n$
(b) By martingale transformation, S_n submartingale $\Rightarrow S_n^+$ is submartingale and by Doob's inequality,
$xP(\max_{0\le m\le n}S_m^+\ge x)\le ES_n^+I\{\max_{0\le m\le n}S_m^+\ge x\}$
(c). $E(\max_{0\le m\le n}S_m^+)^2\le4\sum_{i=1}^n EX_i^2$
$E(\max_{0\le m\le n}S_m^+)^2\le4E(S_n^+)^2$ by L^p maximal inequality.

Theorem 3.2.4 (Continuous Mapping Theorem). Let $g:\mathbb{R}\to\mathbb{R}$ be a measurable function and let $D_g:=\{x:g$ is discontinuous at $x\}$. If $X_n\Rightarrow X$ and $P(X\in D_g)=0$, then

$$g(X_n)\Rightarrow g(X).$$

Proof. By Skorokhod's theorem, we can construct $Y_n\to Y$ a.s. and Y_n (Y resp.) has the same distribution as X_n (X resp.) Because $P(Y\in D_g)=P(X\in D_g)=0$, $g(Y_n)\to g(Y)$ a.s.. This implies $g(Y_n)\Rightarrow g(Y)$; hence $g(X_n)\Rightarrow g(X)$. □

Theorem 3.2.5 (Portmanteau Theorem). The following are equivalent:
(i) $X_n\Rightarrow X_\infty$.
(ii) $\forall$ open sets G, $\liminf_{n\to\infty}P(X_n\in G)\ge P(X_\infty\in G)$.
(iii) $\forall$ closed sets K, $\limsup_{n\to\infty}P(X_n\in K)\le P(X_\infty\in K)$.
(iv) $\forall\ A$ with $P(X_\infty\in\partial A)=0$, $\lim_{n\to\infty}P(X_n\in A)=P(X_\infty\in A)$.

Proof. "(i)$\Longrightarrow$(ii)": By Skorokhod's theorem, we can construct $Y_n\to Y_\infty$ a.s. and Y_n (Y_∞ resp.) has the same distribution as X_n (X_∞ resp.) We have

$$\begin{aligned}\liminf_{n\to\infty}P(X_n\in G)=&\liminf_{n\to\infty}P(Y_n\in G)=\liminf_{n\to\infty}E1_{\{Y_n\in G\}}\\\ge&E\liminf_{n\to\infty}1_{\{Y_n\in G\}} &&\text{(Fatou)}\\\ge&E1_{\{Y_\infty\in G\}} &&\text{(by }Y_n\to Y_\infty\text{ a.s. and }G\text{ is open)}\\=&P(Y_\infty=G)=P(X_\infty\in G).\end{aligned}$$

"(ii)$\Longleftrightarrow$(iii)": Consider $K=G^c$.
"(ii)+(iii)$\Longrightarrow$(iv)": Use $A^o\subset A\subset\bar{A}$.
"(iv)$\Longrightarrow$(i)": Consider $A=(-\infty,x]$. □

$R(\theta,\delta)=\mathbb{E}_{X\sim p(\theta)}\big(L(\theta,\delta(X))\big)$

Data Reduction

Sufficient:

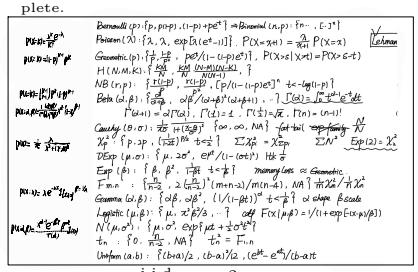
1. def. conditional distribution $[X\mid T{=}t]$ doesn't depend on θ.
2. B&C6.2.2. if $p(x\mid\theta)/q(T(x)\mid\theta)$ is free of θ, then $T(X)$ is suff.
3. NFFC. $T(X)$ is sufficient if.f. $p_\theta(x)=g_\theta(T(x))h(x)$.
   Assuming γ is a counting measure as a countable set $\mathcal{X}$ (i.e. X is discrete) Let the family of pmf's be given by $\{P_\theta:\theta\in\Theta\}$ discrete)
   $\Leftarrow$: Suppose $p_\theta(x)=g_\theta(T(x))h(x)$, $\forall x\in\mathcal{X}$. Need to show that T is sufficient.
   $$P_\theta(X=x\mid T(X)=t)=\frac{P_\theta(X=x,T(X)=t)}{P(T(X)=t)}=$$
   $$\begin{cases}\frac{0}{P(X)\neq t} & T(x)\neq t\\ \frac{P(T(X)=t)}{P(X)\neq t} & T(x)=t\end{cases}$$
   $$=\begin{cases}0 & T(x)\neq t\\ \frac{g_\theta(t)h(x)}{\sum_{y\in x:T(y)=t}g_\theta(t)h(y)} & T(x)=t\end{cases}$$
   $$=\begin{cases}0 & T(x)\neq t\\ \frac{h(x)}{\sum_{y\in x:T(y)=t}h(y)}\perp\theta & T(x)=t\end{cases}$$
   $\Rightarrow$: Suppose is sufficient for θ, so $P_\theta(X=x)=P_\theta(X=x,T(X)=t)$
   $$=P_\theta(X=x\mid T(X)=t)P_\theta(T(X)=t)\triangleq h(x)g_\theta(t)$$
   as $P_\theta(X=x\mid T(X)=t)$ is free of θ by definition

Minimal sufficient:

1. def. sufficient T is min.suff. if T is function of any other suff. T'.
2. if $p(x;\theta){=}c_{x,y}p(y;\theta)\Leftrightarrow T(x){=}T(y)$, then T is min.suff.
   We first prove that T is sufficient. Start with
   $$T(\mathcal{X})=\{t:t=T(x)\text{ for some }x\in\mathcal{X}\}$$
   $$=\text{range of }T.$$
   For each $t\in T(\mathcal{X})$, we consider the preimage $A_t=\{x:T(x)=t\}$ and select an arbitrary representative x_t from each A_t. Then, for any $y\in\mathcal{X}$, we have $y\in A_{T(y)}$ and $X_{T(y)}\in A_{T(y)}$. By the definition of A_t, this implies that $T(y){=}T(X_{T(y)})$. From the assumption of the theorem,
   $$p(y;\theta){=}c_{y,x_{T(y)}}p(x_{T(y)};\theta){=}h(y)g_\theta(T(y))$$
   which yields sufficiency of T by the NFFC.
   Consider another sufficient statistic T'. By NFFC,
   $$p(x;\theta){=}\tilde{g}_\theta(T'(x))\tilde{h}(x)$$
   Take any x,y such that $T'(x){=}T'(y)$, then
   $$p(x;\theta)=\tilde{g}_\theta(T'(x))\tilde{h}(x)=\tilde{g}_\theta(T'(y))\tilde{h}(y)\cdot\frac{\tilde{h}(x)}{\tilde{h}(y)}\overset{4.}{=}p(y;\theta)C_{x,y}$$
   Hence, $T(x){=}T(y)$ by the assumption of the theorem. So $T'(x){=}T'(y)$ implies $T(x){=}T(y)$ for any sufficient statistic T' and any x and y. As a result, T is a minimal sufficient statistic.

Complete: (complete suff. $\Rightarrow$ minimal suff.)

1. def. V is ancillary if the distribution of V is free of θ.
2. def. T is complete if $\mathbb{E}_\theta[f(T)]{=}0$ for all θ implies $f(T){=}0$ a.e.
   Consider $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}\text{CauchyLoc}(\theta)$. Then $A(X){=}X_{(n)}{-}X_{(1)}$ is ancillary even though $(X_1,\cdots,X_n)$ is minimal sufficient. To see this, note that $X_i{=}Z_i{+}\theta$ for $Z_i\overset{\text{i.i.d.}}{\sim}\text{CauchyLoc}(\theta)$, we can see that $X_{(i)}{=}Z_{(i)}{+}\theta$ and $A(X){=}A(Z)\perp\!\!\!\perp\theta$.
3. def. V is first-order ancillary if $\mathbb{E}_\theta[V]$ is free of θ.
   Let $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}\text{Bernoulli}(\theta)$, $\theta\in(0,1)$. Then $T(X){=}\sum_{i=1}^n X_i$ is sufficient.
   Suppose $\mathbb{E}_\theta[f(T(X))]{=}0$ for all $\theta\in(0,1)$,
   $$\sum_{j=0}^n f(j)\binom{n}{j}\theta^j(1-\theta)^{n-j}{=}0,\quad\forall\theta\in(0,1)$$
   Dividing both sides by $(1-\theta)^n$ and reparameterizing $\beta{=}\frac{\theta}{1-\theta}$ we can rewrite it as
   $$\sum_{j=0}^n f(j)\binom{n}{j}\beta^j{=}0,\quad\forall\beta{>}0$$
   if f is non-zero, then LHS is a polynomial of degree at most n. However, an nth-degree polynomial has at most n roots. Hence, it is impossible for the LHS to be equal to 0 for every $\beta{>}0$ unless $f{=}0$. Therefore, T is complete.



Let $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\theta,\sigma^2)$ with unknown θ and an known $\sigma^2{>}0$. Is $\bar{X}_n{=}n^{-1}\sum_{i=1}^n X_i$ complete for this model?
Consider the special case of $n{=}1$ and $\sigma{=}1$. $T(X){=}X{\sim}\mathcal{N}(\theta,1)$
$$\mathbb{E}_\theta[f(X)]{=}0,\quad\forall\theta\in\mathbb{R}$$
Multiplying both sides by $\sqrt{2\pi}\mathrm{e}^{\frac{\theta^2}{2}}$, we have
$$\int_{-\infty}^{\infty}f(x)\exp\Big(-\frac{x^2}{2}\Big)\exp(\theta x)\mathrm{d}x{=}0,\quad\forall\theta\in\mathbb{R}.$$
We decompose f into its positive and negative parts as $f(x){=}f_+(x){-}f_-(x)$, where $f_+(x){=}\max(f(x),0)$, and $f_-(x){=}\max(-f(x),0)$. Note that $f_+{\geq}0$ and $f_-{\geq}0$. For all $x\in\mathbb{R}$ $f_+(x){=}f_(x)$ if and only if $f_+(x){=}f_-(x){=}0$.
Suppose f_+ and f_- have non-zero components, and we may write
$$\frac{\int_{-\infty}^{\infty}f_+(x)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{e}^{\theta x}\mathrm{d}x}{\int_{-\infty}^{\infty}f_+(x)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x}=\frac{\int_{-\infty}^{\infty}f_-(x)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{e}^{\theta x}\mathrm{d}x}{\int_{-\infty}^{\infty}f_-(x)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x}$$
Note that
$$\frac{f_+(x)\mathrm{e}^{-\frac{x^2}{2}}}{\int_{-\infty}^{\infty}f_+(x)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x}$$
defines a probability density.
The equality of the mgfs implies equality of the densities, which in turn implies $f_+(x){=}f_-(x)$ a.e..Then $f_+(x){=}f_-(x){=}0$ a.e., or in other words, $f(x){=}0$ a.e.. Hence T is complete.
3. (Basu) complete and sufficient $U\perp\!\!\!\perp$ ancillary V.
   Example: Let $X_1,..,X_n\overset{i.i.d}{\sim}N(\mu,\sigma^2)$ where both μ,σ^2 are unknown. Then, $\bar{X}_n\perp n^{-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2$ where $\bar{X}_n{=}n^{-1}\sum_{i=1}^N X_i$
   Fix any $\sigma{>}0$ and consider sub-model $P_\sigma{=}\{N(\mu,\sigma^2)\}$. In each sub-model, $\bar{X}_n$ is complete and sufficient, and $n^{-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2$ is ancillary. By Basu's theorem, $\bar{X}_n\perp\sum_{i=1}^n(X_i-\bar{X}_n)^2$ under $N(\mu,\sigma^2)$ for any $\mu\in\mathbb{R}$. Since σ is arbitrary, the conclusion holds for the full model $P{=}\{N(\mu,\sigma^2){:}\mu\in\mathbb{R},\sigma{>}0\}$.
5. (Rao-Blackwell) for conv.loss and suff T, $R(\theta,\mathbb{E}[\delta\mid T])\leq R(\theta,\delta)$. $\mathbb{E}(\mathbb{E}_\theta(L(g(\theta),\delta(x)|T)))\geq\mathbb{E}_\theta(L(g(\theta),\eta(T)))$
6. common steps:
   (a) suppose $\int f(x)h(x)\mathrm{e}^{\theta x}\,dx{=}0$ for all $\theta\in\Omega$,
   (b) decompose $f{=}f_+{-}f_-$ with $f_+,f_-{\geq}0$,
   (c) view f_+ and f_- as un-normalised densities p_+ and p_-,
   (d) argue that MGF of p_+ and p_- are equal, then $f_+{=}f_-$ a.e.

UMRUE: [Example] $X\sim P_{oi}(\theta)$, $g(\theta)=e^{-a\theta}$.
$\mathbb{E}_\theta(\delta(x))=\sum_{x=0}^{\infty}\delta(x)\frac{e^{-\theta}\theta^x}{x!}=e^{-a\theta}$
$\sum_{x=0}^{\infty}\frac{\delta(x)\theta^x}{x!}=e^{(1-a)\theta}=\sum_{x=0}^{\infty}\frac{(1-a)^x\theta^x}{x!}$
Hence $\delta(x)=(1-a)^x$.

1. def. $R(\theta,\delta){\leq}R(\theta,\delta')$ for $\forall\,\theta{\in}\Omega$ and $\forall$ unbiased δ'.
2. (Lehmann-Scheffe) if T is comp.suff. and $\mathbb{E}_\theta[h(T)]{=}g(\theta)$, then $h(T)$ is i) only unbiased fun. of T, ii) UMRUE under conv.loss. [unique UMVUE under strictly conv. loss]
3. δ_0 is UMVUE of $g(\theta)$ if.f. $\mathbb{E}[\delta_0(X)U]{=}0$ for all U with mean 0. (Hint: consider $\delta_\lambda{=}\delta_0{+}\lambda U$.) to prove $\delta_1(x)+\delta_2(x)$ UMVUE.
4. strategies to find UMRUE
   (a) Rao-B. condition of comp.suff. on unbiased estimator.
   cond.probability for discrete, ancillary for continuous.
   (b) Solve the unique δ satisfying $\mathbb{E}_\theta[\delta(T)]{=}g(\theta)$.
   (c) Guess.
   Example (Rao-Blackwellisation) Suppose $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}$ Bernoulli (θ). We know that $T(X){=}\sum_{i=1}^n X_i$ is a complete sufficient statistic, we also know that $n^{-1}T(X)$ is an unbiased estimator for θ, i.e.
   $$\mathbb{E}_\theta\Big(\frac{T(X)}{n}\Big){=}\frac{1}{n}\mathbb{E}_\theta\Big(\sum_{i=1}^n X_i\Big){=}\frac{n\theta}{n}{=}\theta.$$
   So $n^{-1}T(X)$ is an UMRUE for θ under any convex loss function. Suppose we want to estimate θ^2, let's examine: $\delta(X){=}\mathbb{I}\{X_1{=}X_2{=}1\}{=}X_1{\cdot}X_2$. Then we have
   $$\mathbb{E}_\theta[\delta(X)]=\mathbb{E}_\theta[\mathbb{I}\{X_1=X_2=1\}]$$
   $$=\mathbb{E}_\theta\left(X_1\cdot X_2\right)\overset{\text{i.i.d.}}{=}\left[\mathbb{E}_\theta\left(X_1\right)\right]^2=\theta^2.$$
   By conditioning on $T(X)$, we can find the UMRUE via
   $$\mathbb{E}_\theta[\delta(X)\mid T(X)=t]=\mathbb{P}_\theta\left(X_1=X_2=1\mid T(X)=t\right)$$
   $$=\frac{\mathbb{P}_\theta\left(X_1=X_2=1,\sum_{i=1}^n X_i=t-2\right)}{\mathbb{P}_\theta(T(X)=t)}$$
   $$=\frac{\theta^2\binom{n-2}{t-2}\theta^{t-2}(1-\theta)^{n-t}\mathbb{I}\{t\geq2\}}{\binom{n}{t}\theta^2(1-\theta)^{n-t}}$$
   $$=\frac{t(t-1)\mathbb{I}\{t\geq2\}}{n(n-1)}.$$
   Note that in this case $\mathbb{I}\{t{\geq}2\}$ is redundant as for $t{=}0$ or 1 , the term $t(t{-}1){=}0$. Hence, we conclude that the UMRUE for θ^2 is
   $$\frac{T(X)(T(X)-1)}{n(n-1)}.$$
   Example (Rao-Blackwellisation) Suppose $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}U(0,\theta)$. In this case, $T(X){=}X_{(n)}$ is a complete and sufficient statistic, and $\delta(X){=}2X$ is an unbiased estimator for θ, i.e. $\mathbb{E}_\theta(2X){=}2\times\frac{\theta}{2}{=}\theta$.
   Given $X_{(n)}$,X_1 is equal to $X_{(n)}$ with probability $1/n$ and follows uniform $\left(0,X_{(n)}\right)$ with probability $1{-}1/n$. Hence,
   $$\mathbb{P}_\theta\left(X_1=x_1\mid T(X)\right)=\frac{1}{n}\times\mathbb{I}\{T(X)=x_1\}$$
   $$+\left(1-\frac{1}{n}\right)\times\frac{1}{T(X)}\mathbb{I}\{0<x_1<T(X)\}$$
   To find the UMVUE, we calculate
   $$\mathbb{E}_\theta[\delta(X)\mid T(X)]=2\mathbb{E}_\theta[X_n\mid T(X)]=$$
   $$2\left\{\frac{1}{n}T(X)+\left(1-\frac{1}{n}\right)\int_0^{T(X)}\frac{x_1}{T(X)}dx_1\right\}$$
   $$=2\left\{\frac{T(X)}{n}+\left(1-\frac{1}{n}\right)\frac{T(X)}{2}\right\}$$
   $$=\frac{n+1}{n}T(X).$$
   Example (Guess) Let $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}N\left(\mu,\sigma^2\right)$. Consider the case where $\theta{=}\left(\mu,\sigma^2\right)$ is unknown.
   1) The UMVUE for σ^2 is $\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2{=}S^2$
   2) How about the UMVUE for σ?
   3) What is the UMVUE for μ^2?
   For 2), Observe that
   $$X_i-\bar{X}_n\sim N\Big(0,\frac{n-1}{n}\sigma^2\Big)\quad\Rightarrow$$
   $$\mathbb{E}\{|X_i-\bar{X}_n|\}=\sigma\sqrt{\tfrac{2}{\pi}}\times\sqrt{\tfrac{n-1}{n}}$$
   This implies
   $$\delta'=\frac{\sqrt{\pi n}}{\sqrt{2(n-1)}}|X_i-\bar{X}_n|$$
   is unbiased for σ. While the Rao-Blackwellisation is very tedious. Instead, we can observe another fact that
   $$S_*^2{=}\textstyle\sum_{i=1}^n(X_i-\bar{X}_n)^2{=}(n{-}1)S^2$$
   And $S_*^2{\sim}\sigma^2\chi^2_{n-1}$. Hence, $\mathbb{E}(S_*){=}\sigma\mathbb{E}\left(\chi_{n-1}\right)$, which implies that
   $$\frac{\mathbb{E}(S_*)}{\mathbb{E}\left(\chi_{n-1}\right)}{=}\sigma$$
   meaning that $\frac{S_*}{\mathbb{E}\left(\chi_{n-1}\right)}$ is unbiased for σ and hence UMVU.
   For 3), Taking the expecatation of the UMVUE for μ and squaring it, we obtain
   $$\mathbb{E}\left(\bar{X}_n^2\right){=}\mu^2{+}\frac{\sigma^2}{n}$$
   So
   $$\delta_n(X){=}\bar{X}_n^2-\frac{S_*^2}{n(n-1)}$$
   is the UMVUE. However, $\delta_n(X)$ can be negative even though the estimand is a non-negative quantity. The estimator is in fact inadmissible and dominated by the biased estimator $\max(0,\delta_n(X))$ since
   $$\mathbb{E}[\max(0,\delta_n(X))]{\neq}\max\{\mathbb{E}(\delta(X)),0\}.$$
   Example 5 (Semiparametric unbiased estimators). For $n{>}2$, let $X_1,\ldots,X_n\overset{\text{iid}}{\sim}F$, where F is some unknown distribution on $\mathbb{R}$. Suppose that F is symmetric about some unknown point $\theta\in\mathbb{R}$. That is, suppose for all $X{\sim}F$, we have
   $$X{=}_d2\theta{-}X.$$
   Consider the model
   $\mathcal{F}{=}\{$ all distributions on $\mathbb{R}$ with finite variance symmetric about $\theta\}$. Then there is no UMVUE for the point of symmetry θ. Proof. Suppose for sake of contradiction that the UMVUE $T(X)$ exists. Since $\bar{X}$ is unbiased for the full model $\mathcal{F}$,$T(X)$ must have variance no larger than $\bar{X}$. However, we know that $\bar{X}$ is the unique UMVUE for the Gaussian submodel, $\{\mathcal{N}(\theta,1){:}\ \theta\in\mathbb{R}\}$, and so $T(X)$ must equal $\bar{X}$ a.s. in the Gaussian submodel. This implies that $T(X){=}\bar{X}$ a.s. under any continuous distribution on $\mathbb{R}$. In particular, $T(X){=}\bar{X}$ a.s. under the uniform submodel $\{\text{Unif}(\theta{-}1,\theta{+}1){:}\ \theta\in\mathbb{R}\}$. However, we learned in Homework 2, Problem 3 (b), that the distinct estimator $\frac{1}{2}\left(X_{(1)}{+}X_{(n)}\right)$ has strictly better variance than $\bar{X}$ in the uniform submodel. Since $\frac{1}{2}\left(X_{(1)}{+}X_{(n)}\right)$ is also unbiased for the full model $\mathcal{F}$,$T(X)$ cannot be the UMVUE after all. We are forced to conclude that no UMVUE exists over the whole family.
   $X\sim U(\theta-1,\theta+1)$ UMVUE

Fisher Information:

1. $I(\theta){=}\mathbb{E}[(\frac{\partial}{\partial\theta}\log f_\theta(x))^2]{=}{-}\mathbb{E}[\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)]$.
2. Cramer-Rao lower bound: $\text{Var}(\delta){\geq}[g'(\theta)]^2/I(\theta)$.
   $\varphi(x){=}\frac{\partial}{\partial\theta}\log f_\theta(x),\mathbb{E}_\theta[\varphi(X)]{=}0,E_\theta[\delta^2]{<}\infty,g'(\theta){=}\mathbb{E}_\theta[\delta\varphi]$.
3. If $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}P_\theta$, $I_X(\theta){=}I_{X_1}(\theta{+}\cdots{+}I_{X_n}(\theta)){=}nI_{X_1}(\theta)$.
   Let $\mathcal{P}$ be a one-parameter exponential family in canonical form and density p_η given by
   $$p_\eta{=}exp(\eta T(x){-}A(\eta))h(x),$$
   then
   $$\frac{\partial\log p_\eta(x)}{\partial\eta}{=}T(x){-}A'(\eta).$$
   By the previous results, we have
   $I(\eta){=}Var_\eta(T(x){-}A'(\eta)){=}Var_\eta(T(x)){=}A''(\eta)$.
   If the family is parameterised instead by $\mu{=}A'(\eta){=}\eta\mathbb{T}(\mathbb{X})$, then
   $$A''(\eta){=}I(\mu)(A''(\eta))^2,$$
   and so, because $A''(\eta){=}Var(T)$, we have $I(\mu){=}\frac{1}{Var_\eta(T)}$. Observe also that because T is UMVUE for μ, the lower bound variance $Var(\delta){\geq}(I(\mu))^{-1}$ for unbiased estimation δ of μ is sharp in this example.
   Suppose X us ab absolutely continuous random variable with density f. The family of distributions $\mathcal{P}{=}\{P_\theta{:}\theta\in\mathbb{R}\}$ with P_θ the distribution of $\theta{+}\epsilon$ is called a location family.
   $$\int g(x)\mathrm{d}P_\theta(x)=\mathbb{E}\theta(g(X))$$
   $$=\mathbb{E}_\theta(g(\theta+\epsilon))$$
   $$=\int g(\theta+\epsilon)f(\epsilon)\mathrm{d}\epsilon$$
   $$=\int g(x)f(x-\theta)\mathrm{d}x.$$
   So P_θ has the density $p_\theta{=}f(x{-}\theta)$. The corresponding Fisher information for the family is
   $$I(\theta)=\mathbb{E}_\theta(\frac{\partial\log f(X-\theta)}{\partial\theta})^2$$
   $$=\mathbb{E}_\theta(-\frac{f'(X-\theta)}{f(X-\theta)})^2$$
   $$=\mathbb{E}(\frac{f'(\epsilon)}{f(\epsilon)})^2$$
   $$=\int\frac{\left\{f'(x)\right\}^2}{f(x)}\mathrm{d}x\perp\!\!\!\perp\theta.$$
   So, for the location families, $I(\theta)$ is constant with respect to θ.

Suppose $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}N(\mu,\sigma^2),\mu\in\mathbb{R}$, and $\sigma^2{>}0$. We want to estimate $g_1(\mu,\sigma^2){=}\mu$ and $g_2(\mu,\sigma^2){=}\sigma^2$. Look at only unbiased estimators.

Claim 1: $\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X_n})^2$ is UMVUE for σ^2, where $\overline{X_n}=\frac{1}{n}\sum_{i=1}^n X_i$.

Proof. (a) $(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2)$ is complete and sufficient.
(b) $\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X_n})^2$ is a function of $(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2)$. This is because $\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X_n})=\sum_{i=1}^n X_i^2-n\overline{X_n}^2$.
(c) $\frac{1}{\sigma^2}\sim\chi^2_{n-1}\Rightarrow\mathbb{E}(\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X_n})^2){=}\frac{\sigma^2}{n-1}(n{-}1){=}\sigma^2$ □

Claim 2: $\overline{X_n}$ is UMVNE for μ. $\mathbb{E}(\overline{X_n}-\mu)^2{=}\sigma^2/n$. Note that$\mathbb{E}(\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X_n})^2/(n-1)-\sigma^2){=}Var(\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X_n})^2/(n-1)){=}\frac{2(n-1)\sigma^4}{(n-1)^2}{=}\frac{2\sigma^4}{(n-1)}$.

$$\log p_{\mu,\sigma^2}(X){=}-\tfrac{n}{2}\log\sigma^2-\textstyle\sum_{i=1}^n\frac{(X_i-\mu)^2}{2\sigma^2}+C_n,$$
$$\frac{\partial\log p_\theta(X)}{\partial\mu}{=}\textstyle\sum_{i=1}^n\frac{X_i-\mu}{\sigma^2},$$
$$\frac{\partial^2\log p_\theta(X)}{\partial\mu^2}{=}-\frac{n}{\sigma^2},$$
$$\frac{\partial\log p_\theta(X)}{\partial\sigma^2}{=}-\frac{n}{\sigma^2}{+}\textstyle\sum_{i=1}^n\frac{(X_i-\mu)^2}{2\sigma^4}$$
$$\frac{\partial^2\log p_\theta(X)}{(\partial\sigma^2)^2}{=}\frac{n}{\sigma^4}-\textstyle\sum_{i=1}^n\frac{(X_i-\mu)^2}{\sigma^6}$$
$$\frac{\partial^2\log p_\theta(X)}{\partial\mu\partial\sigma^2}{=}-\textstyle\sum_{i=1}^n\frac{X_i-\mu}{\sigma^4}$$
Therefore, the Fisher information matrix is
$$I{=}\begin{pmatrix}\frac{n}{\sigma^2}&0\\0&\frac{n}{2\sigma^4}\end{pmatrix}$$
$$I_{22}{=}-\tfrac{n}{2\sigma^4}{+}\mathbb{E}(\textstyle\sum_{i=1}^n\frac{(X_i-\mu)^2}{\sigma^6})=-\frac{n}{2\sigma^4}+\frac{n\sigma^2}{\sigma^6}=\frac{n}{2\sigma^4}$$
$\Rightarrow$ CRLB for $\mu{=}(1\quad0)I^{-1}\binom{1}{0}{=}\frac{1}{I_{11}}{=}\frac{\sigma^2}{n}$.
$\Rightarrow$ CRLB for $\sigma^2{=}(0\quad1)I^{-1}\binom{0}{1}{=}\frac{1}{I_{22}}{=}\frac{2\sigma^4}{n}$.

Exponential Family

general form:
$p(x;\theta)=\exp\{\sum_{i=1}^n \eta_i(\theta)T_i(x)-B(\theta)\}h(x)$.

1. standardiser: $B(\theta)=\log\int\exp\{\sum_{i=1}^n \eta_i(\theta)T_i(x)\}h(x)\,dx$.
2. parameter space: $\Theta=\{\theta:B(\theta)<\infty\}$.
3. Moments Take $f(x)=1$, then

$$\frac{\partial A(\eta)}{\partial \eta_i}=\int T_i(x)\exp\left(\sum_{j=1}^s \eta_j T_j(x)-A(\eta)\right)h(x)\mathrm{d}\mu(x)$$
$$=\mathbb{E}_\eta[T_i(X)]$$
$$\frac{\partial^2 A(\eta)}{\partial\eta_i\partial\eta_j}=\mathrm{Cov}_\eta\left(T_i(X),T_j(X)\right)$$

canonical form:
$p(x;\eta)=\exp\{\sum_{i=1}^n \eta_i T_i(x)-A(\eta)\}h(x)$.

1. natural parameter η_i, and nature parameter space.
2. def. canonical exp.fam. is minimal, if no affine T_i's and η_i's.
   ($\sum_i \lambda_i T_i(x)=\lambda_0$ implies $\lambda_i=0$, similar for η_i's.)
3. def. min.exp.fam. is full-rank, if nat.par.space contain open rect.
4. if exp.fam. is full-rank, then T is minimal sufficient and complete.

Hypothesis Testing

Basic definition:

1. test function $\phi(x)$: the prob. rejects H_0 given $X=x$,
2. power function: $\beta(\theta)=\mathbb{E}_\theta[\phi(X)]=P_\theta(\text{rejects } H_0)$,
3. significant level α: $\sup_{\theta\in\Omega_0}\mathbb{E}_\theta[\phi(X)]\le\alpha$,
4. level-α uniformly most powerful test ϕ: if $\mathbb{E}_\theta[\phi(X)]\ge\mathbb{E}_\theta[\phi^*(X)]$ for all $\theta\in\Omega_1$, for any other level-α test ϕ^*.
5. families with monotone likelihood ratio in $T(X)$:
   (a) $\theta\neq\theta'$ implies $p_\theta\neq p_{\theta'}$, (identifiability)
   (b) $\theta<\theta'$ implies the ratio $p_\theta(x)/p_{\theta'}(x)$ is a non-decreasing function of $T(X)$. (monotonicity)

Proof: For $\alpha=0$ and $\alpha=1$ the theorem is easily seen to be true provided the value $k=+\infty$ is admitted in (3.8) and $0\cdot\infty$ is interpreted as 0 . Throughout the proof we shall therefore assume $0<\alpha<1$.

(i): Let $\alpha(c)=P_0\{p_1(X)>cp_0(X)\}$. Since the probability is computed under P_0, the inequality need be considered only for the set where $p_0(x)>0$, so that $\alpha(c)$ is the probability that the random variable $p_1(X)/p_0(X)$ exceeds c. Thus $1-\alpha(c)$ is a cumulative distribution function, and $\alpha(c)$ is nonincreasing and continuous on the right, $\alpha(c-0)-\alpha(c)=P_0\{p_1(X)/p_0(X)=c\}$,$\alpha(-\infty)=1$, and $\alpha(\infty)=0$. Given any $0<\alpha<1$, let c_0 be such that $\alpha(c_0)\le\alpha\le\alpha(c_0-0)$, and consider the test ϕ defined by

$$\phi(x)=\begin{cases}1 & \text{when } p_1(x)>c_0p_0(x)\\ \frac{\alpha-\alpha(c_0)}{\alpha(c_0-0)-\alpha(c_0)} & \text{when } p_1(x)=c_0p_0(x)\\ 0 & \text{when } p_1(x)<c_0p_0(x)\end{cases}$$

Here the middle expression is meaningful unless $\alpha(c_0)=\alpha(c_0-0)$; since then $P_0\{p_1(X)=c_0p_0(X)\}=0$,$\phi$ is defined a.e. The size of ϕ is

$$E_0\phi(X)=P_0\left\{\frac{p_1(X)}{p_0(X)}>c_0\right\}+\frac{\alpha-\alpha(c_0)}{\alpha(c_0-0)-\alpha(c_0)}P_0\left\{\frac{p_1(X)}{p_0(X)}=c_0\right\}$$

so that c_0 can be taken as the k of the theorem.

(ii): Suppose that ϕ is a test satisfying (3.7) and (3.8) and that ϕ^* is any other test with $E_0\phi^*(X)\le\alpha$. Denote by $S^+$and $S^-$the sets in the sample space where $\phi(x)-\phi^*(x)>0$ and <0 respectively. If x is in S^+,$\phi(x)$ must be >0 and $p_1(x)\ge kp_0(x)$. In the same way $p_1(x)\le kp_0(x)$ for all x in S^-, and hence $\int(\phi-\phi^*)(p_1-kp_0)d\mu=\int_{S^+\cup S^-}(\phi-\phi^*)(p_1-kp_0)d\mu\ge0$.

The difference in power between ϕ and $\phi*$ therefore satisfies

$$\int(\phi-\phi^*)p_1d\mu\ge k\int(\phi-\phi^*)p_0d\mu\ge0$$

as was to be proved.

(iii): Let ϕ^* be most powerful at level α for testing p_0 against p_1, and let ϕ satisfy (3.7) and (3.8). Let S be the intersection of the set $S^+\cup S^-$, on which ϕ and ϕ^* differ, with the set $\{x:p_1(x)\neq kp_0(x)\}$, and suppose that $\mu(S)>0$. Since $(\phi-\phi^*)(p_1-kp_0)$ is positive on S, it follows from Problem 2.4 that $\int_{S^+\cup S^-}(\phi-\phi^*)(p_1-kp_0)d\mu=\int_S(\phi-\phi^*)(p_1-kp_0)d\mu>0$ and hence that ϕ is more powerful against p_1 than ϕ^*. This is a contradiction, and therefore $\mu(S)=0$, as was to be proved.

If ϕ^* were of size $<\alpha$ and power <1, it would be possible to include in the rejection region additional points or portions of points and thereby to increase the power until either the power is 1 or the size is α. Thus either $E_0\phi^*(X)=\alpha$ or $E_1\phi^*(X)=1$.

Example:

Let $X_1,\ldots,X_n \overset{\text{ii.d}}{\sim}\mathcal{N}(\mu,\sigma^2)$, with σ known. Consider the following two hypotheses:

$$H_0{:}\mu=0 \text{ and } H_1{:}\mu=\mu_1$$

where μ_1 is known. Since this is a simple case (only two distributions), we calculate the likelihood ratio:

$$r(x)=\frac{p_1(x)}{p_0(x)}=\frac{\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-(x_i-\mu_1)^2/2\sigma^2\right)}{\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-x_i^2/2\sigma^2\right)}$$
$$=\exp\left(\frac{1}{\sigma^2}\mu_1\sum_{i=1}^n x_i-\frac{n\mu_1^2}{2\sigma^2}\right).$$

We observe that the likelihood ratio is only a function of the sufficient statistic (this is in general true by the Factorization Criterion). We have the following equivalences:

$$r(x)>k\Leftrightarrow\mu_1\frac{\sum x_i}{\sigma^2}-\frac{n\mu_1^2}{2\sigma^2}>\log k$$
$$\Leftrightarrow\mu_1\sum x_i>k'$$
$$\Leftrightarrow\begin{cases}\sum x_i>k'' & \text{if } \mu_1>0\\ \sum x_i<k''' & \text{if } \mu_1<0\end{cases}$$

Let us focus on the first case where $\mu_1>0$ (but it is important to note that the two cases $\mu_1>0$ versus $\mu_1<0$ induce different rejection regions). We can rewrite the test in a different form so that the left hand side of the inequality has standard normal distribution under the null:

$$\Leftrightarrow\frac{\sqrt{n}\bar{x}}{\sigma}>k''''$$

Note: Here the sufficient statistic essentially determines a MP test. Also, observe that for a given level α, the constant involved in the LRT is uniquely determined by the constraint $\mathbb{E}_{p0}\phi(X)=\alpha$ and thus depends only on the distribution of the null hypothesis and not at all on μ_1 (provided that $\mu_1>0$).

Thus we have by the NP lemma that the test: reject H_0 iff $\sqrt{n}\bar{x}/\sigma>k(\alpha)$ is MP where we pick $k'''=k(\alpha)$ (where $k(\alpha)\equiv z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal) so that it equals the target level. This value is uniquely determined by the size constraint:

$$\mathbb{E}_{p0}\phi(X)=\alpha=\mathbb{P}_{\mu=0}\left(\frac{\sqrt{n}\bar{X}}{\sigma}>k^{m'''}\right).$$

Now an important observation: for any $\mu_1>0$ the MP test is the same, which means that it is actually a UMP at level α for testing:

$$H_0{:}\mu=0 \text{ and } H_1{:}\mu>0.$$

So we see no μ_1 dependence here and it's really nice that I can test against any $\mu_1>0$ at once. Similarly, we could derive a distinct UMP test for testing against $H_1{:}\mu<0$. Unfortunately, no UMP test exists for testing

$$H_0{:}\mu=0 \text{ and } H_1{:}\mu\neq0$$

because the $\mu>0$ test dominates the $\mu<0$ test in the $\mu>0$ scenario and vice versa i.e. the shape of the most powerful rejection rejection (whether we reject for $\bar{x}>k$ or whether we reject for $\bar{x}<k$) is dependent on the sign of μ_1. **Example:** Suppose X has a binomial distribution with success probability θ and $n=2$ trials. If we are interested in testing $H_0{:}\theta=1/2$ versus $H_1{:}\theta=2/3$, then

$$L(X)=\frac{p_1(X)}{p_0(X)}=\frac{\binom{2}{X}(2/3)^X(1/3)^{2-X}}{\binom{2}{X}(1/2)^X(1/2)^{2-X}}=\frac{2^X\times4}{9}.$$

Suppose the desired significance level is $\alpha=50\%$. Let $\phi(x_1,x_2)=\begin{cases}1 & \text{if } (x_1,x_2)=(1,1)\\ 0 & \text{if } (x_1,x_2)=(0,0)\end{cases}$, randomised when we observe (0,1) and (1,0), such that $E_{\theta_0}(\phi(X_1,X_2))=\frac{1}{2}$. **Example**(One parameter exponential family). Consider the case where $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)\propto h(x)\exp(\theta T(x))$, and we are interested in testing

$$H_0{:}\theta=\theta_0 \quad \text{vs.} \quad H_1{:}\theta=\theta_1.$$

We want an MP test at level α. The likelihood ratio is

$$\frac{\prod_i p_{\theta_1}(x_i)}{\prod_i p_{\theta_0}(x_i)}\propto\exp((\theta_1-\theta_0)\sum_i T(x_i)).$$

Since the exponential is just a monotone transformation, an MP test will reject for large $(\theta_1-\theta_0)\sum_i T(x_i)$. Assuming $\theta_1>\theta_0$, we will reject for large $\sum_i T(x_i)$. That is, an MP test has the form

$$\phi(x)=\begin{cases}1 & \text{if } \sum_i T(x_i)>k\\ \gamma & \text{if } \sum_i T(x_i)=k\\ 0 & \text{if } \sum_i T(x_i)<k\end{cases}$$

where k,γ are chosen to satisfy the size constraint
$\alpha=\mathbb{E}_{\theta_0}\phi(X)=\mathbb{P}_{\theta_0}[\sum_i T(X_i)>k]+\gamma\mathbb{P}_{\theta_0}[\sum_i T(X_i)=k]$.

Note that $\sum_i T(x_i)$ has no explicit θ dependence and that k,γ do not depend on θ_1 (assuming $\theta_1>\theta_0$). This means ϕ is in fact UMP for testing

$$H_0{:}\theta=\theta_0 \quad \text{vs.} \quad H_1{:}\theta>\theta_0.$$

Here, H_1 is an example of a one-sided alternative, which arises when the parameter values of interest lie on only one side of the real-valued parameter θ_0.

Corrollary 3.2.1 (TSH): Let $\beta=\beta(\theta_1)$ denote the power of the Most Powerful Test for testing H_0: $\theta=\theta_0$ and $H_1{:}\theta=\theta_1$ at level $\alpha\in(0,1)$. Then $\beta\ge\alpha$. Furthermore, $\beta>\alpha$ unless $P_{\theta_1}=P_{\theta_0}$

Proof: Let ϕ be the Most Powerful Test from (i) of Neyman-Pearson lemma

Let $\psi(x)\equiv\alpha\Rightarrow\beta=E_{\theta_1}(\phi(x))\ge E_{\theta_1}(\psi(x))=\alpha$ suppose $\alpha=\beta$, then $\psi(x)$ is a Most Powerful test

$$\Rightarrow P_{\theta_1}(x)=kP_{\theta_0}(x) \text{ as } \mu_1\Rightarrow k=1$$
$$\Leftrightarrow P_{\theta_0}=P_{\theta_1}$$

Find UMP test:

1. Neyman-Pearson Lemma for simple vs simple:
   Existence there exist $\phi(x)$ and constant k
   (a) $\mathbb{E}_{p0}[\phi(X)]=\alpha$, (size = level)
   (b) $\phi(x)=\begin{cases}1 & \text{if } \frac{p_1(x)}{p_0(x)}>k_\alpha\\ 0 & \text{if } \frac{p_1(x)}{p_0(x)}<k_\alpha\end{cases}$
   Sufficiency if a test holds (a) and (b) for some k, then it is MP.
   Necessity if a test ϕ is MP at level α, then it holds (b) for some k, and also holds (a) unless $\exists$ a test of size $<\alpha$ and power 1.
2. For MLR family, test $H_0{:}\theta\le\theta_0$ vs $H_1{:}\theta\ge\theta_1$:
   (a) there exist a UMP test at level α of the form
   $$\phi(x)=\begin{cases}1 & \text{if } T(x)>k\\ \gamma & \text{if } T(x)=k\\ 0 & \text{if } T(x)<k\end{cases},$$
   (b) the power function $\beta(\theta)=\mathbb{E}_\theta[\phi(X)]$ is strictly increasing for $0<\beta(\theta)<1$.

Optimal Tests for Composite Nulls:

Hypothesis: $H_0{:}X\sim f_\theta,\ \theta\in\Omega$ vs $H_1{:}X\sim g$.

Consider new hypothesis H_Λ: $X\sim h_\Lambda(x)=\int_{\Omega_0}f_\theta(x)\,d\Lambda(\theta)$.

Let β_Λ be the power of MP level-α test ϕ_Λ for H_Λ vs H_1.

Prior Λ is a least favorable dist if $\beta_\Lambda\le\beta_{\Lambda'}$ for any prior Λ'.

Suppose ϕ_Λ is an MP level α-test for testing H_Λ against H_1. If ϕ_Λ is level-α for the original hypothesis H_0, then

1. The test ϕ_Λ is MP for original test vs alternative,
2. The prior distribution Λ is least favorable.

Example 2 (Double exponential). Let $X\sim$DoubleExponential(θ), with density $p_\theta(x)=\frac{1}{2}\exp(-|x-\theta|)$. It is easy to see that the model is identifiable, so we need only check the second condition. Fix any $\theta'>\theta$ and consider the likelihood ratio

$$\frac{p_{\theta'}(x)}{p_\theta(x)}=\exp\left(|x-\theta|-\left|x-\theta'\right|\right).$$

Note that

$$|x-\theta|-\left|x-\theta'\right|=\begin{cases}\theta-\theta' & \text{if } x<\theta\\ 2x-\theta-\theta' & \text{if } \theta\le x\le\theta'\\ \theta'-\theta & \text{if } x>\theta',\end{cases}$$

which is non-decreasing in x. Therefore, the family has MLR in $T(x)=x$. Finally, we give an example of a model that does not exhibit MLR in $T(x)=x$.

Example 3 (Cauchy location model). Let X have density $p_\theta(x)=\frac{1}{\pi}\cdot\frac{1}{1+(x-\theta)^2}$. We find two points for which the MLR condition fails. For any fixed $\theta>0$,

$$\frac{p_\theta(x)}{p_0(x)}=\frac{1+x^2}{1+(x-\theta)^2}\to1 \quad \text{as } x\to\infty \text{ or } x\to-\infty,$$

but $p_\theta(0)/p_0(0)=1/\left(1+\theta^2\right)$, which is strictly less than 1 . Thus the ratio must increase at some values of x and decrease at others. In particular, it is not monotone in x. Here we have shown that the likelihood ratio in $T(x)=x$ is not MLR.

Theorem 1 (TSH 3.8.1). Suppose ϕ_Λ is a MP level- α test for testing H_Λ against g. If ϕ_Λ is level- α for the original hypothesis H_0 (i.e., $\mathbb{E}_{\theta_0}\phi_\Lambda(x)\le\alpha,\forall\theta_0\in\Omega_0$), then 1. The test ϕ_Λ is MP for original $H_0{:}\theta\in\Omega_0$ vs. g. 2. The distribution Λ is least favorable.

Proof. 1. Let ϕ^* be any other level- α test of $H_0{:}\theta\in\Omega_0$vs.g. Then ϕ^* is also a level- α test for H_A vs. g, because
$\mathbb{E}_\theta\phi^*(X)=\int\phi^*(x)f_\theta(x)d\mu(x)\le\alpha,\forall\theta\in\Omega_0$,
which implies that
$\int\phi^*(x)h_\Lambda(x)d\mu(x)=\iint\phi^*(x)f_\theta(x)d\mu(x)d\Lambda(\theta)\le\int\alpha d\Lambda(\theta)=\alpha$.

Since ϕ_A is MP for H_A vs. g, we have
$\int\phi^*(x)g(x)d\mu(x)\le\int\phi_\Lambda(x)g(x)d\mu(x)$,

Hence ϕ_A is a MP test for H_0 vs. g, because ϕ_A is also level α. 2. Let Λ' be any distribution on Ω_0. Since $\mathbb{E}_\theta\phi_\Lambda(x)\le\alpha,\forall\theta\in\Omega_0$, we know that ϕ_Λ must be level α for $H_{\Lambda'}$ vs. g. Thus $\beta_\Lambda\le\beta_{\Lambda'}$, so Λ is the least favorable distribution.

Example 1 (Testing in the presence of nuisance parameters). Let $X_1,\ldots,X_n$ be i.i.d. $\mathcal{N}(\theta,\sigma^2)$, where both θ,σ^2 are unknown. We consider testing $H_0{:}\sigma\le\sigma_0$ against $H_1{:}\sigma>\sigma_0$. To find a UMP test, we follow the previously mentioned strategy: 1. First we fix a simple alternative (θ_1,σ_1) for some arbitrary θ_1 and $\sigma_1>\sigma_0$. 2. Second, we choose a prior distribution Λ to collapse our null hypothesis over. Intuitively, the least favorable prior should make the alternative hypothesis hard to distinguish. Hence, a rule of thumb consists in concentrating Λ on the boundary between H_1 and H_0 (i.e. the line $\{\sigma=\sigma_0\}$). Thus Λ will be a probability distribution over $\theta\in\mathbb{R}$ for the fixed $\sigma=\sigma_0$. Another useful observation is that, given any test function $\phi(x)$ and a sufficient statistic T, there exists a test function η that has the same power as ϕ but depends on x only through T :
$\eta(T(x))=\mathbb{E}[\phi(x)|T(x)]$.

Hence, we can restrict our attention to the sufficient statistics (Y,U), where $Y=\bar{X}$ and $U=\sum_{i=1}^n(X_i-X)^2$. We know that $Y\sim\mathcal{N}\left(\theta,\sigma^2/n\right),U\sim\sigma^2\chi^2_{n-1}$, and Y is independent of U by Basu's theorem. Thus, for Λ supported on $\sigma=\sigma_0$, we obtain the joint density of (Y,U) under H_Λ as

$$c_0u^{\frac{n-3}{2}}\exp\left(-\frac{u}{2\sigma_0^2}\right)\int\exp\left(-\frac{n}{2\sigma_0^2}(y-\theta)^2\right)d\Lambda(\theta)$$

and the joint density under alternative hypothesis (θ_1,σ_1) as

$$c_1u^{\frac{n-3}{2}}\exp\left(-\frac{u}{2\sigma_1^2}\right)\exp\left(-\frac{n}{2\sigma_1^2}(y-\theta_1)^2\right).$$

From the above observations, we see that the choice of Λ only affects the distribution of Y. To achieve minimal maximum power against the alternative (i.e., to be least favorable), we need to choose Λ such that the two distributions become as close as possible. Under the alternative hypothesis, $Y\sim\mathcal{N}\left(\theta_1,\frac{\sigma_1^2}{n}\right)$. Under H_Λ, the distribution of Y is in a convolution form, i.e., $Y=Z+\Theta$ for $Z\sim\mathcal{N}\left(0,\frac{\sigma_0^2}{n}\right),\Theta\sim\Lambda$, where Z and Θ are independent. Hence, if we choose $\Theta\sim\mathcal{N}\left(\theta_1,\frac{\sigma_2^2-\sigma_0^2}{n}\right),Y$ will have the same distribution under the null and the alternative, which is $\mathcal{N}\left(\theta_1,\frac{\sigma_1^2}{n}\right)$. Under this choice of prior, the LRT rejects for large values of $\exp\left(-\frac{u}{2\sigma_1^2}+\frac{u}{2\sigma_0^2}Z\right)$, i.e., it rejects for large values of u (since $\sigma_1>\sigma_0$). So the MP test rejects H_Λ if $\sum_{i=1}^n(X_i-\bar{X})^2$ lies above some threshold determined by the size constraint. In particular, it rejects if $\sum_{i=1}^n(X_i-\bar{X})^2>\sigma_0^2C_{n-1,1-\alpha}$, where $C_{n-1,1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of χ^2_{n-1}. 3. Next we check if the MP test is level- α for the composite null. For any (θ,σ) with $\sigma\le\sigma_0$, the probability of rejection is:

$$\mathbb{P}_{\theta,\sigma}\left(\frac{\sum_{i=1}^n(X_i-X)^2}{\sigma^2}>\frac{\sigma_0^2C_{n-1,1-\alpha}}{\sigma^2}\right)=$$
$$\mathbb{P}\left(\chi^2_{n-1}>\frac{\sigma_0^2}{\sigma^2}C_{n-1,1-\alpha}\right)\le\alpha,$$

while equality holds iff $\sigma=\sigma_0$. Hence, it follows from Theorem 1 that our test is MP for testing the original null H_0vs.$\mathcal{N}(\theta_1,\sigma_1)$. 4. Finally, the MP level- α test for testing the composite null H_0 vs. the arbitrarily chosen (θ_1,σ_1) does not depend on the choice of (θ_1,σ_1). Hence it is UMP for testing the original composite null vs. the composite alternative.

Example 2 (Nonparametric Quality Checking). Identical light bulbs have lifetime $X_1,\ldots,X_n$ with an arbitrary distribution $\mathbb{P}$ over $\mathbb{R}$. Let u be a fixed threshold for a satisfactory lifetime and $\mathbb{P}(X\le u)$ be the probability of a given light bulb being unsatisfactory. Given the data of sample lifetimes we may be interested in testing whether the probability of having an unsatisfactory light bulb is too large:

$$H_0{:}\mathbb{P}(X\le u)\ge p_0 \text{ vs. } H_1{:}\mathbb{P}(X\le u)<p_0.$$

Here p_0 is a fixed quality parameter. 0. Before we start our search for the UMP test, let us reparametrize the distribution $\mathbb{P}$ as follows. Let $\mathbb{P}^-$and $\mathbb{P}^+$be the conditional distributions of $X|X\le u$ and $X|X>u$ respectively, and let $p=\mathbb{P}(X\le u)$. Then, $\mathbb{P}$ has a one-to-one correspondence with $\left(\mathbb{P}^+,\mathbb{P}^-,p\right)$. For any fixed $\mathbb{P}$, let $p^-$and p^+be the conditional densities of $\mathbb{P}^-$and $\mathbb{P}^+$ with respect to some measure μ (existence of the densities and base measure can be justified, e.g. by Radon-Nikodym theorem in measure theory). The joint density of $X_1,\ldots,X_n$ at values

$x_1,\ldots,x_n$ when $x_{i_1},\ldots,x_{i_m}\le u<x_{j_1},\ldots,x_{j_{n-m}}$ is then given by
$p^m\left(\prod_{j=1}^m p^-\left(x_{i_j}\right)\right)(1-p)^{n-m}\left(\prod_{k=1}^{n-m} p^+\left(x_{j_k}\right)\right)$.
1. As before, we fix a simple alternative $\left(\mathbb{P}^-,\mathbb{P}^+,p_1\right)$ where $p_1<p_0$. 2. We next choose a proper prior. We guess that Λ mostly concentrates on the boundary point $\left(p^+,p^-,p_0\right)$. If so, for testing H_Λ vs. the simple alternative, the LRT rejects for large values of
$$\frac{\binom{n}{m}p_1^m(1-p_1)^{n-m}}{\binom{n}{m}p_0^m(1-p_0)^{n-m}}$$
which is equivalent to testing $\mathrm{Bin}(n,p_0)$ vs. $\mathrm{Bin}(n,p_1)$. Thus, the MP test, which rejects for small values of $m=\#\{i:X_i\le u\}$[1], is given by
$$\phi_\Lambda(x)=\begin{cases}1, & \text{if } m<k\\ \gamma, & \text{if } m=k\\ 0, & \text{if } m>k\end{cases}$$
where k and γ are both determined by the level constraint $\mathbb{E}_{p0}\phi_\Lambda(x)=\alpha$. 3. Now we check if ϕ_Λ is level- α for our composite null H_0. Note that the power function of ϕ_Λ depends on $\mathbb{P}$ only through $p=\mathbb{P}(X\le u)$. Given that this family has MLR in m, the power function would be monotone. So for any $p>p_0$, the rejection probability under the null is still smaller than α. Hence, ϕ_Λ is the MP test for testing the composite null H_0 against the simple alternative $H_1:\left(\mathbb{P}^-,\mathbb{P}^+,p_1\right)$. 4. Finally, ϕ_Λ has no dependence on the choice of alternative hypothesis. Therefore, ϕ_Λ is UMP for testing the composite null H_0 against the composite alternative H_1.

Bayes Esti. and Average Risk Optimality

Basic definition:

1. loss function: $L(\theta,d)$,
2. risk: $R(\theta,\delta)=\mathbb{E}_\theta[L(\theta,\delta(X))]$,
3. average/Bayes risk: $r(\Lambda,\delta)=\int_\Omega R(\theta,\delta)\,d\Lambda(\theta)$,
4. posterior risk: $\mathbb{E}[L(\Theta,\delta(X))\mid X=x]$,
5. Bayes estimator δ_Λ: δ that minimizes $r(\Lambda,\delta)$,
6. Bayes risk: $r(\Lambda,\delta_\Lambda)$ for prior Λ,
7. inadmissible (esimator δ): if there exists another estimator δ' which dominates it (that is, such that $R(\theta,\delta')\le R(\theta,\delta)$ for all θ, with strict inequality for some θ),
8. admissible: no such dominating estimator δ' exists.

1. Example $L(\theta,d)=w(\theta)(d-g(\theta))^2$
Here the function $w(\theta)$ can be thought of as a weight function. Then the Bayes estimator minimizes the posterior risk
$$\mathbb{E}\left[w(\Theta)(g(\Theta)-d)^2\mid X=x\right]$$
with respect to d,
$d^2\mathbb{E}[w(\Theta)|X=x]-2d\mathbb{E}[w(\Theta)g(\Theta)|X=x]+\mathbb{E}\left[w(\Theta)g^2(\Theta)|X=x\right]$
Taking derivatives w.r.t. d, we see that this function takes on its minimum when
$2d\mathbb{E}[w(\Theta)|X=x]-2\mathbb{E}[w(\Theta)g(\Theta)|X=x]=0$.
the Bayes estimator is
$$\delta_\Lambda(X)=\frac{\mathbb{E}[w(\Theta)g(\Theta)|X=x]}{\mathbb{E}[w(\Theta)|X=x]}.$$
2. Beta × Bin
$$\pi(\theta\mid x)\propto\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$
$$\propto\theta^{x+a-1}(1-\theta)^{n-x+b-1}\sim\mathrm{Beta}(x+a,n-x+b)$$
$$\mathbb{E}[\Theta|X=x]=\frac{x+a}{n+a+b}$$
$$\frac{X+a}{n+a+b}=\frac{n}{n+a+b}\left(\frac{X}{n}\right)+\frac{a+b}{n+a+b}\left(\frac{a}{a+b}\right).$$
3. Normal mean estimation $\mathrm{N}\left(\Theta,\sigma^2\right)$, with σ^2 known.
Let $X_1,\ldots,X_n\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\Theta,\sigma^2\right)$, with σ^2 known. Further, let $\Theta\sim\mathcal{N}\left(\mu,b^2\right)$, where μ, and b are fixed prior hyper-parameters.
Thus, we have that, the posterior $\pi(\theta|x)$ is
$$\pi(\theta\mid x)\propto\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right)(b)$$
$$\propto\exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)\theta^2+\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)\theta\right).$$
$$\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)=\frac{\text{mean}}{\text{variance}},$$
$$-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)=-\frac{1}{2\cdot\text{ variance}}.$$
So the posterior distribution is $\mathcal{N}\left(\tilde{\mu},\tilde{\sigma}^2\right)$ with
$$\tilde{\mu}=\frac{n\bar{x}/\sigma^2+\mu/b^2}{n/\sigma^2+1/b^2},\quad\tilde{\sigma}^2=\frac{1}{n/\sigma^2+1/b^2}.$$
Under squared error loss, the Bayes estimator is the posterior mean, which can be written as
$$\tilde{\mu}=\frac{n/\sigma^2}{n/\sigma^2+1/b^2}\bar{x}+\frac{1/b^2}{n/\sigma^2+1/b^2}\mu.$$

Properties:

1. THM(TPE4.2.3) unbiased estimator δ for $g(\theta)$ is not Bayes est. under the squared loss function unless its average risk is zero.
$$\mathbb{E}\left[(\delta(X)-g(\Theta))^2\right]=0,$$
where the expectation is taken over X and Θ. Proof. Let δ be an unbiased Bayes estimator under squared error loss satisfying the assumptions of the theorem. Then, we know that δ is the mean of the posterior distribution, i.e.,
$$\delta(X)=\mathbb{E}[g(\Theta)|X]\text{ a.s.}$$
Thus we have that,
$$\mathbb{E}[\delta(X)g(\Theta)]=\mathbb{E}[\mathbb{E}[\delta(X)g(\Theta)|X]]=\mathbb{E}[\delta(X)\mathbb{E}[g(\Theta)|X]]=\mathbb{E}[\delta^2(X)],$$
where (a) follows by substituting for $\mathbb{E}[g(\Theta)|X]]$, using (8.1). We also have that,
$$\mathbb{E}[\delta(X)g(\Theta)]=\mathbb{E}[\mathbb{E}[\delta(X)g(\Theta)|\Theta]]=\mathbb{E}[g(\Theta)\mathbb{E}[\delta(X)|\Theta]]\overset{(b)}{=}\mathbb{E}\left[g^2(\Theta)\right],$$
where (b) follows since $\mathbb{E}[\delta(X)|\Theta]=g(\Theta)$ because δ is an unbiased estimator of $g(\theta)$. We have that the average risk under squared error,
$$\mathbb{E}\left[(\delta(X)-g(\Theta))^2\right]=\mathbb{E}\left[\delta^2(X)\right]-2\mathbb{E}[\delta(X)g(\Theta)]+\mathbb{E}\left[g^2(\Theta)\right],$$
$$=\mathbb{E}\left[\delta^2(X)\right]-\mathbb{E}[\delta(X)g(\Theta)]+\mathbb{E}\left[g^2(\Theta)\right]-\mathbb{E}[\delta(X)g(\Theta)],$$
$$\overset{(c)}{=}\mathbb{E}\left[\delta^2(X)\right]-\mathbb{E}\left[\delta^2(X)\right]+\mathbb{E}\left[g^2(\Theta)\right]-\mathbb{E}\left[g^2(\Theta)\right],$$
$$=0,$$
where (c) follows, by substituting for $\mathbb{E}[\delta(X)g(\Theta)]$ once from (8.2) and once from (8.3). Thus, we have that $\mathbb{E}\left[(\delta(X)-g(\Theta))^2\right]=0$, i.e., the average risk is zero, proving the claim.
2. Example
Let $X_1,\ldots,X_n\sim\mathcal{N}\left(\theta,\sigma^2\right)$, with $\sigma^2>0$ known. Is $\bar{X}$ Bayes under squared error for some choice of prior distribution? We know that, $\mathbb{E}(\bar{X}|\theta)=\theta$, i.e., $\bar{X}$ is an unbiased estimator of θ. Further, we have that the average risk under squared error,
$$\mathbb{E}\left[(\bar{X}-\Theta)^2\right]=\frac{\sigma^2}{n}\neq0,$$
which means that $\bar{X}$ is not the Bayes estimator under any prior distribution!
3. THM. (TPE 5.2.4) Any unique Bayes estimator is admissible. Proof. Suppose δ_Λ is Bayes for Λ, and for some δ', $R\left(\theta,\delta'\right)\le R(\theta,\delta_\Lambda)$ for all $\theta\in\Omega$. If we take expectations with respect to θ, the inequality is preserved, and we get
$\int_{\theta\in\Omega}R\left(\theta,\delta'\right)d\Lambda(\theta)\le\int_{\theta\in\Omega}R(\theta,\delta_\Lambda)d\Lambda(\theta)$
This implies that δ' is also Bayes since δ' has risk less than or equal to δ_Λ, which minimizes the average risk, and thus $\delta'=\delta_\Lambda$ with probability 1 for all P_θ.
4. Bayes estimator is unique when
   (a) under strictly convex loss function,
   (b) $r(\Lambda,\delta)<\infty$, finiteness for comparison,
   (c) $P_\theta<<Q$, where Q is the marginal dist. of X.
   (open support of Λ, and $P_\theta(A)$ cont. w.r.t. θ)
   (fitness for comparison, same support)
5. all admissible estimators are limits of Bayes estimators.

Empirical Bayes estimator:

1. calculate marginal distribution of X:
$m(x\mid r)=\int f(x\mid\theta)\pi(\theta\mid r)\,d\theta$
2. estimate the hyperparameter based on max $m(x\mid r)$.
3. minimize the empirical posterior loss:
$\min_\delta\int L(\theta,\delta(x))\pi(\theta\mid x,\hat{r}(x))d\theta$.

Minimaxity & Worst-Case Optimality

Basic Definition:

1. minimax estimator: δ that minimize $\sup_{\theta\in\Omega}R(\theta,\delta)$.
2. least favorable prior Λ: $r_\Lambda\ge r_{\Lambda'}$ for any prior Λ'.
3. least favourable sequence of priors $\{r_{\Lambda_m}\}$:
   (a) $r_{\Lambda_m}=r(\Lambda_m,\delta_{\Lambda_m})\to r<\infty$,
   (b) $r\ge r_{\Lambda'}$ for any prior Λ'.

Find minimax estimator:

1. If Bayes risk = minimax risk, i.e. $r_\Lambda=\sup_\theta R(\theta,\delta_\Lambda)$,
   (a) Bayes estimator δ_Λ is minimax,
   (b) Λ is a least favourable prior,
   (c) unique Bayes esti. implies unique minimax esti.
2. (TPE5.1.5) If a Bayes estimator has constant risk, it's minimax.
3. (TPE5.1.6) $\omega_\Lambda=\{\theta:R(\theta,\delta_\Lambda)=\sup_{\theta'}R(\theta',\delta_\Lambda)\}$, δ_Λ is minimax if $\Lambda(\omega_\Lambda)=1$.
4. (TPE5.1.2)If a sequence of priors $\{r_{\Lambda_m}\}$ with $r_{\Lambda_m}\to r<\infty$, and there exists estimator δ with $\sup_\theta R(\theta,\delta)=r$, then
   (a) δ is minimax,
   (b) $\{r_{\Lambda_m}\}$ is least-favourable.
5. minimax for Binom(n,θ)
$X\sim\mathrm{Binom}(n,\theta)$ for some $\theta\in(0,1)$ and that we use the squared error loss function. the sample proportion $\frac{X}{n}$ NOT minimax? The risk of this estimator is
$$R\left(\theta,\frac{X}{n}\right)=\frac{\theta(1-\theta)}{n}.$$
The risk has a unique maximum at $\theta=\frac{1}{2}$, so the worst-case risk is
$$\sup_{\theta\in\Omega}R\left(\theta,\frac{X}{n}\right)=R\left(\frac{1}{2},\frac{X}{n}\right)=\frac{1}{4n}.$$
Unfortunately, we cannot apply TPE5.1.6 directly because if $\Lambda\left(\left\{\frac{1}{2}\right\}\right)=1$, then $\delta_\Lambda(X)=\frac{1}{2}\neq\frac{X}{n}$.
However, we can use the Corollary 5.1.5 to find a minimax estimator and then compare the risk of the minimax estimator with that of $\frac{X}{n}$. To find a minimax estimator, we will search for a prior such that the Bayes estimator has constant risk.
Recall the following useful fact. Under the prior distribution $\mathrm{Beta}(a,b)$, the Bayes estimator under the squared error loss is
$$\delta_{a,b}(X)=\frac{X+a}{n+a+b}.$$
For any a and b,
$$R\left(\theta,\delta_{a,b}\right)=\mathbb{E}_\theta\left[\left(\frac{X+a}{n+a+b}-\theta\right)^2\right]$$
$$=\frac{1}{(n+a+b)^2}\left(n\theta(1-\theta)+(a(\theta-1)+\theta b)^2\right).$$
This is a quadratic function of θ. To eliminate the θ dependence in $R\left(\theta,\delta_{a,b}\right)$, we need the coefficients of the linear and quadratic terms to equal zero. The coefficient of θ^2 is
$$-n+(a+b)^2,$$
so we need $a+b=\sqrt{n}$ (since $a,b>0$). The coefficient of θ is
$$n-2a(a+b)=n-2a\sqrt{n},$$
so we need $a=b=\frac{\sqrt{n}}{2}$. With these choices of a and b, the risk of $R\left(\theta,\delta_{a,b}\right)$ is constant, which implies that Beta $\left(\frac{\sqrt{n}}{2},\frac{\sqrt{n}}{2}\right)$ is a least favorable prior with constant risk. Then our Bayes estimator
$$\delta_{\frac{\sqrt{n}}{2},\frac{\sqrt{n}}{2}}(X)=\frac{X+\frac{\sqrt{n}}{2}}{n+\sqrt{n}},$$
is minimax with constant risk of
$$\frac{1}{4(\sqrt{n}+1)^2}.$$
Since the worst-case risk of $\frac{X}{n}$ is $\frac{1}{4n}>\frac{1}{4(\sqrt{n}+1)^2}$, we can conclude that $\frac{X}{n}$ is not minimax.
6. minimax for N$(\theta,\sigma^2$ konwn$\sigma)$
Let $X_1,\ldots,X_n\overset{iid}{\sim}\mathcal{N}\left(\theta,\sigma^2\right)$, with σ^2 known. Our goal is to estimate θ under squared-error loss. pick the natural estimator $\bar{X}$. Note that it has constant risk $\frac{\sigma^2}{n}$, which suggests minimaxity because we know that Bayes estimators with constant risk are also minimax estimators. However, $\bar{X}$ is not Bayes for any prior, because under squared-error loss unbiased estimators are Bayes estimators only in the degenerate situations of zero risk (TPE Theorem 4.2.3), and $\bar{X}$ is unbiased. Thus, we cannot conclude by our previous results (e.g., TPE Corollary 5.1.5) that $\bar{X}$ is minimax.
We might try to consider the wider class of estimators $\delta_{a,\mu_0}(X)=a\bar{X}+(1-a)\mu_0$ for $a\in(0,1)$ and $\mu_0\in\mathbb{R}$, because many of the Bayes estimators we've encountered are convex combinations of a prior and a data mean. Note however that the worst case risk for these estimators is infinite:
$$\sup_\theta\mathbb{E}_\theta[\theta-\delta(X)]^2=\sup_\theta\left\{a^2\,\mathrm{Var}_\theta(\bar{X})+(1-a)^2(\theta-\mu_0)^2\right\}$$
$$=\frac{a^2\sigma^2}{n}+(1-a)^2\sup_\theta(\theta-\mu_0)^2$$
$$=+\infty.$$
Since these estimators have poorer worst case risk than $\bar{X}$, they certainly cannot be minimax. We could keep trying to find Bayes estimators with better worst-case performance than $\bar{X}$, but we would fail: it turns out that $\bar{X}$ is in fact minimax. To establish this, we will extend our minimax results to the limits of Bayes estimators, it suffices to find a sequence $\{\Lambda_m\}$ such that $r_{\Lambda_m}\to\frac{\sigma^2}{n}=:r$. Using the conjugate prior is a good starting point, so we let $\{\Lambda_m\}$ be the conjugate priors $\left\{N\left(0,m^2\right)\right\}$ with variance tending to ∞, so that Λ_m tends to the (improper with $\pi(\theta)=1,\forall\theta\in\mathbb{R}$) uniform prior on $\mathbb{R}$. By TPE Example 4.2.2, the posterior for θ associated with each Λ_m is
$$\theta|X_1,\ldots,X_n\sim N\left(\frac{\frac{n\bar{X}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{m^2}},\frac{1}{\frac{n}{\sigma^2}+\frac{1}{m^2}}\right).$$
In particular, the posterior variance does not depend on $X_1,\ldots,X_n$, so Lemma 1 below automatically yields the Bayes risk
$$r_{\Lambda_m}=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{m^2}}\xrightarrow{m\to\infty}\frac{\sigma^2}{n}=\sup_\theta R(\theta,\bar{X}).$$
It follows from Theorem 1 that $\bar{X}$ is minimax and $\{\Lambda_m\}$ is least favorable.
Minimax for i.i.d. Normal random variables, unknown mean and variance
let $X_1,\ldots,X_n\overset{iid}{\sim}\mathcal{N}\left(\theta,\sigma^2\right)$, with both θ and σ^2 unknown. Note that
$$\sup_{\theta,\sigma^2}R\left(\left(\theta,\sigma^2\right),\bar{X}\right)=\sup_{\sigma^2}\frac{\sigma^2}{n}=\infty$$
and in fact, the maximum risk of any estimator in this setting is infinite, so the question of minimaxity is uninteresting. Therefore, we restrict attention to the family parameterized by $\Omega=\left\{\left(\theta,\sigma^2\right):\theta\in\mathbb{R},\sigma^2\le B\right\}$, where B is a known constant. Assume δ is any other estimator. Calculating the risk of $\bar{X}$ within this family, we find
$$\sup_{\theta\in\mathbb{R},\sigma^2\le B}R\left(\left(\theta,\sigma^2\right),\bar{X}\right)=\frac{B}{n}$$
$$=\sup_{\theta\in\mathbb{R},\sigma^2=B}R\left(\left(\theta,\sigma^2\right),\bar{X}\right)$$
$$\le\sup_{\theta\in\mathbb{R},\sigma^2=B}R\left(\left(\theta,\sigma^2\right),\delta\right)\text{ [submodel minimax]}$$
$$\le\sup_{\theta\in\mathbb{R},\sigma^2\le B}R\left(\left(\theta,\sigma^2\right),\delta\right)$$
where the first inequality follows from the fact that $\bar{X}$ is minimax for i.i.d. normals with known σ^2, and the second inequality follows from the fact that we are taking the supremum over a larger set. Hence, we are able to show that $\bar{X}$ is minimax over Ω by focusing on the case where σ^2 is known. Notice further that the form of the estimator does not depend on the upper bound B, though the bound is necessary for minimaxity to be worth investigating.
8. Minimax for binomial random variables, weighted squared error loss. Let $X\sim\mathrm{Bin}(n,\theta)$ with the loss function $L(\theta,d)=\frac{(d-\theta)^2}{\theta(1-\theta)}$. This is a simple weighted squarederror loss with the weights $w(\theta)=\frac{1}{\theta(1-\theta)}$ it penalizes errors near 0 and 1 more strongly than errors near $\frac{1}{2}$.
Note that for any $\theta,R\left(\theta,\frac{X}{n}\right)=\frac{1}{n}$; that is, the risk is constant in θ, suggesting $\frac{X}{n}$ is minimax. We will show that this is indeed the case. We should be careful since TPE Theorem 4.2.3 is only valid under the squared-error loss. Since our loss function is different, an unbiased estimator can be Bayes. In this example, this is indeed the case.
Recall from TPE Corollary 4.1.2 that the Bayes estimator associated with the loss $L(d,\theta)=w(\theta)(d-\theta)^2$ is given by $\frac{\mathbb{E}_{\Theta|X}[\Theta w(\Theta)|X]}{\mathbb{E}_{\Theta|X}[w(\Theta)|X]}$. Invoking this result, we find that the Bayes estimator has the form
$$\delta_\Lambda(X)=\frac{\mathbb{E}_{\Theta|X}\left[\frac{1}{1-\Theta}\middle|X\right]}{\mathbb{E}_{\Theta|X}\left[\frac{1}{\Theta(1-\Theta)}\middle|X\right]}.$$
This is true for arbitrary priors Λ, but to calculate a closed form Bayes estimator, we use a prior conjugate to the binomial likelihood: $\Theta\sim\Lambda_{a,b}=\mathrm{Beta}(a,b)$, for some $a,b>0$.
Suppose we observe $X=x$. If $a+x>1$ and $b+n+x>1$, then substituting the result of Remark 1 below into equation 10.1 proves that the estimator
$$\delta_{a,b}(x)=\frac{a+x-1}{a+b+n-2},$$
minimizes the posterior risk. In particular, the estimator $\delta_{1,1}(x)=\frac{x}{n}$ minimizes the posterior risk with respect to the uniform prior after observing $0<x<n$. If we can verify that this form remains unchanged when $x\in\{0,n\}$, then the estimator $\delta_{1,1}(X)=\frac{X}{n}$ is Bayes with constant risk, and hence minimax.
To see that this is the case, note that the posterior risk under the prior $\Lambda_{1,1}$ after observing $X=x$ and deciding $\delta(x)=d$ is
$\int_0^1\frac{(d-\theta)^2}{\theta(1-\theta)}\cdot\frac{\Gamma(x+1+n-x+1)}{\Gamma(x+1)\Gamma(n-x+1)}\cdot\theta^x(1-\theta)^{n-x}d\theta$,
which, in the case $X=0$, simplifies to
$$\int_0^1\frac{(d-\theta)^2}{\theta}(1-\theta)^{n-1}d\theta$$
This integral converges for $d=0$ and diverges otherwise, so the posterior risk is minimized by choosing $\delta(0)=0$. Similarly, in the case $X=n$, the posterior risk is minimized by choosing $\delta(n)=1=\frac{n}{n}$. This confirms that $\delta_{1,1}(X)=\frac{X}{n}$ minimizes the posterior risk for any outcome X, and is indeed Bayes. Since as we mentioned before this estimator has constant risk we can conclude that $\frac{X}{n}$ is minimax.
Notice that the form of the minimax estimator here depends on the type of loss being used: $\frac{X}{n}$ has constant risk for the type of weighted

$R(\hat{\theta}_n,\theta)=E_\theta\left(\frac{X}{n}-\theta\right)^2=\frac{\theta(1-\theta)}{n}$ $\sup_\theta R(\hat{\theta}_n,\theta)=\frac{1}{4n}$
$R(\hat{\theta}_n,\theta)=\frac{1}{4(\sqrt{n}+1)^2}$ constant
$\lim_{n\to\infty}\sup_\theta R(\hat{\theta}_n,\theta)/\sup_\theta R(\hat{\theta}_n,\theta)=\lim_{n\to\infty}\frac{1}{4n}\Big/\frac{1}{4(\sqrt{n}+1)^2}=1$
$\lim_{n\to\infty}\frac{R(\hat{\theta}_n,\theta)}{R(\hat{\theta}_n,\theta)}=4\theta(1-\theta)<1\quad\forall\theta\neq\frac{1}{2}$

squared error loss considered here.

9. remark1 Remark 1. Recall that the Beta function can be evaluated as
$$\int_0^1 x^{k_1-1}(1-x)^{k_2-1}dx=\frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}$$
whenever $k_1,k_2>0$. Therefore, if $Y\sim$ Beta(a,b), where $a,b>0$, we can explicitly evaluate the expectation
$$\mathbb{E}\left[\frac{1}{1-Y}\right]=\int_0^1\frac{1}{1-y}\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1}\right]dy$$
$$=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(a)\Gamma(b-1)}{\Gamma(a+b-1)}$$
where in the second step we require $b>1$ in order to apply the relation 10.2. A similar argument yields
$$\mathbb{E}\left[\frac{1}{Y(1-Y)}\right]=\frac{(a+b-2)(a+b-1)}{(a-1)(b-1)},$$
whenever $a>1$. Combining these identities, we have that, whenever $a,b>1$,
$$\frac{\mathbb{E}\left[\frac{1}{1-Y}\right]}{\mathbb{E}\left[\frac{1}{Y(1-Y)}\right]}=\frac{a-1}{a+b-2}.$$

10. Randomized Minimax Let $X\sim$Bin(n,θ), where $\theta\in[0,1]$, and consider estimation of θ under the $0-1$ loss,
$$L(\theta,d)=\begin{cases}0 & \text{if } |d-\theta|<\alpha\\ 1 & \text{otherwise}\end{cases}.$$
First consider an arbitrary non-random estimator δ. Since X can take on only the $n+1$ values $\{0,1,\ldots,n\}$, the estimator $\delta(X)$ can take on only $n+1$ values, $\{\delta(0),\delta(1),\ldots,\delta(n)\}$. If $\alpha<\frac{1}{2(n+1)}$, then we can always find θ_0 such that $|\delta(x)-\theta_0|\geq\alpha$ for every $x\in\{0,\ldots,n\}$; see Figure 10.1. Hence, $R(\theta_0,\delta(X))=1$ is the maximum risk of any non-random δ.
Consider instead the estimator $\delta'(X,U)=U$, which is completely random and independent of the data X. Then, for any $\theta\in[0,1]$,
$$R\left(\theta,\delta'\right)=\mathbb{E}\left[L\left(\theta,\delta'(X,U)\right)\right]$$
$$=\mathbb{P}(|U-\theta|\geq\alpha)$$
$$=1-\mathbb{P}(\theta-\alpha<U<\alpha+\theta)$$
$$\leq 1-\alpha<1$$
and since $\alpha>0$, the maximum risk of δ' is smaller than the maximum risk of any non-random δ. Hence, in this setting, there can be no deterministic minimax estimator.

11. Lemma 1 (TPE 5.1.15). Suppose that δ is minimax for a submodel $\theta\in\Omega_0\subset\Omega$ and
$$\sup_{\theta\in\Omega_0}R(\theta,\delta)=\sup_{\theta\in\Omega}R(\theta,\delta)$$
Then, δ is minimax for the full model, $\theta\in\Omega$.

12. Example Let $X_1,\ldots,X_n$ be i.i.d $\mathcal{N}\left(\mu,\sigma^2\right)$, where both μ and σ^2 are unknown. Thus, our parameter vector, $\theta=\left(\mu,\sigma^2\right)$ and our parameter space $\Omega=\mathbb{R}\times\mathbb{R}^+$. Our task now is to estimate μ. Our loss function is the relative squared error loss, given by:
$$L\left(\left(\mu,\sigma^2\right),d\right)=\frac{(d-\mu)^2}{\sigma^2}$$
We consider the submodel where $\sigma^2=1$. That is, $\Omega_0=\mathbb{R}\times\{1\}$, and our loss function simplifies to our usual squared error loss: $L((\mu,1),d)=(d-\mu)^2$. We saw in Example 1 of Lecture 10 that under this loss $\bar{X}$ is minimax for Ω_0. Moreover,
$$R\left(\left(\mu,\sigma^2\right),\bar{X}\right)=\frac{1}{n}\quad\forall\left(\mu,\sigma^2\right)\in\Omega$$
Thus, the risk does not depend on σ^2. Since $R((\mu,1),\bar{X}))=R\left(\left(\mu,\sigma^2\right),\bar{X}\right)\Big)$, we have that the maximum risks are equal. (That is, $\sup_{\theta\in\Omega_0}R(\theta,\delta)=\sup_{\theta\in\Omega}R(\theta,\delta)$. Therefore, it follows from Lemma 1 that $\bar{X}$ is minimax on Ω. Note that, thanks to our new loss function, we don't need to impose boundedness on our variance (like we did in our previous lecture) to establish minimaxity in a meaningful way.

13. Example 2 Suppose $X_1,X_2,\ldots,X_n$ are i.i.d with common CDF F, with mean $\mu(F)<\infty$, and variance $\sigma^2(F)<\infty$. Our goal is to find a minimax estimate of $\mu(F)$ under squared error loss. Constraint (a). Assume $\sigma^2(F)\leq B$.

Now, we've seen in the previous lecture that $\bar{X}$ is minimax for the Gaussian submodel in this case. So a natural guess for us to make is that $\bar{X}$ is minimax. We verify this by application of Lemma 1. First, we compute the supremum risk for the full model:
$$R(F,\bar{X})=\frac{1}{n^2}\sum_i\mathbb{E}(X_i-\mu(F))^2=\frac{\sigma^2(F)}{n}.$$
Since $\sigma^2(F)\in[0,B]$ by assumption, we get:
$$\sup_F R(F,\bar{X})=\frac{B}{n}$$
Now we saw in Lecture 10 that for the submodel $\mathcal{F}_0=\mathcal{N}\left(\mu,\sigma^2\right)$ when $\sigma^2\leq B,\bar{X}$ is minimax. Further, the supremum risk in this case is identical to that of the full model:
$$\sup_{F\in\mathcal{F}_0}R(F,\bar{X})=\frac{B}{n}$$
Thus, using Lemma 1 we conclude that $\bar{X}$ is minimax for the full model. (That is, the nonparametric model still constrained to have $\sigma^2(F)\leq B$.)
Constraint (b). Assume $F\in\mathcal{F}$ where $\mathcal{F}$ is the set of all CDFs with support contained in $[0,1]$. it turns out that $\bar{X}$ isn't minimax.
To show this, first consider the submodel, $\mathcal{F}_0=\{\text{Ber}(\theta)\}_{\theta\in(0,1)}$. Let $Y=\sum_{i=1}^n X_i$ so that $Y\sim\text{Bin}(n,\theta)$ and $\bar{X}=Y/n$. Recall from Lecture 9 that the minimax estimator for $\mu(F)=\theta$, in the Binomial case, is:
$$\delta(X)=\frac{\sqrt{n}}{1+\sqrt{n}}\bar{X}+\frac{1}{2}\left(\frac{1}{1+\sqrt{n}}\right)$$
which has supremum risk $\frac{1}{4(1+\sqrt{n})^2}$. So
$$\sup_\theta R(\theta,\bar{X})=\frac{1}{4n}>\frac{1}{4(1+\sqrt{n})^2}=\sup_\theta R(\theta,\delta)$$
Thus, $\bar{X}$ has a higher worst-case risk than $\delta(X)$ as defined above, and hence, we have shown that $\bar{X}$ is not minimax.
Now, let's get more ambitious, and try to see if we can find the minimax estimator under the full model. We know that this can't be $\bar{X}$, but it's possible that it could be $\delta(X)$. To examine this possibility, we conjecture that $\delta(X)$ is also minimax under the full model. If we are to establish this under the Lemma, we need to show that the supremum risk of $\delta(X)$ under the full model is no more than $\frac{1}{4(1+\sqrt{n})^2}$ (which is the supremum risk for the binomial submodel). Let us compute:
$$\mathbb{E}_F[\delta(X)-\mu(F))^2]$$
$$=\mathbb{E}_F\left[\left(\left(\frac{\sqrt{n}}{1+\sqrt{n}}\right)(\bar{X}-\mu(F))+\frac{1}{1+\sqrt{n}}\left(\frac{1}{2}-\mu(F)\right)\right)^2\right]$$
$$=\left(\frac{1}{1+\sqrt{n}}\right)^2\left[\mathbb{E}\left(X_1^2\right)+\frac{1}{4}-\mu(F)\right]$$
where the third step follows from the fact that $\text{Var}(X_1)=n\,\text{Var}(\bar{X})=\mathbb{E}\left[X_1^2\right]-(\mathbb{E}[X_1])^2=\mathbb{E}\left[X_1^2\right]-(\mu(F))^2$ By assumption $X_1\in[0,1]$, so $X_1^2\leq X_1$ and we can bound the risk:
$$\mathbb{E}_F[\delta(X)-\mu(F))^2]\leq\left(\frac{1}{1+\sqrt{n}}\right)^2\left[\mathbb{E}(X_1)+\frac{1}{4}-\mu(F)\right]$$
$$=\frac{1}{4(1+\sqrt{n})^2}.$$
So, $\delta(X)$ is minimax for the Binomial submodel, and its worst-case risk is the same for the full model and for the Binomial submodel. Therefore, applying the Lemma, we conclude that $\delta(X)$ is minimax. Thus, we have found a minimax estimator.

Property:

1. minimax esti. may not necessarily be Bayes esti.
2. admissible with constant risk, implies minimax.
3. minimaxity may not guarantee admissibility.
4. Example Let $X_1,X_2,\ldots,X_n\overset{iid}{\sim}\mathcal{N}\left(\theta,\sigma^2\right)$ where σ^2 is known, and θ is the estimand. Then the minimax estimator is $\bar{X}$ under squared error loss, and we would like to determine whether $\bar{X}$ is admissible.
Instead of answering this directly, we answer a more general question: when is $a\bar{X}+b$, $a,b\in\mathbb{R}$, (basically, any affine function of $\bar{X}$) admissible? Case 1: $0<a<1$. $a\bar{X}+b$ is a convex combination of $\bar{X}$ and b. it is a Bayes estimator with respect to some Gaussian prior on θ. Further, since we are using squared error loss, which is strictly convex, this Bayes estimator is unique. So, by Theorem 5.2.4, $a\bar{X}+b$ is admissible. Case 2: $a=0$. b is also a unique Bayes estimator with respect to a degenerate prior distribution with unit mass at $\theta=b$. So by Theorem 5.2.4,b is admissible. Case 3: $a=1,b\neq0$. $\bar{X}+b$ is not admissible because it is dominated by $\bar{X}$. $\bar{X}$ has the same variance as $\bar{X}+b$, but strictly smaller bias. In general, the risk of $a\bar{X}+b$ is:
$$\mathbb{E}[(a\bar{X}+b-\theta)]^2=\mathbb{E}\left[(a(\bar{X}-\theta)+b+\theta(a-1))^2\right]$$
$$=\frac{a^2\sigma^2}{n}+(b+\theta(a-1))^2$$
Case 4: $a>1$.
$$\mathbb{E}\left[(a\bar{X}+b-\theta)^2\right]\geq\frac{a^2\sigma^2}{n}>\frac{\sigma^2}{n}=R(\theta,\bar{X}).$$
$\bar{X}$ dominates $a\bar{X}+b$ when $a>1$, and so in this case $a\bar{X}+b$ is inadmissible. Case 5: $a<0$.
$$\mathbb{E}\left[(a\bar{X}+b-\theta)^2\right]>(b+\theta(a-1))^2$$
$$=(a-1)^2\left(\theta+\frac{b}{a-1}\right)^2$$
$$>\left(\theta+\frac{b}{a-1}\right)^2,$$
and this is the risk of predicting the constant $-b/(a-1)$. So, $-b/(a-1)$ dominates $a\bar{X}+b$, and therefore, $a\bar{X}+b$ is again inadmissible.
Case 6: $a=1,b=0$. Here, we use a limiting Bayes argument. Suppose $\bar{X}$ is inadmissible. Then, assuming w.l.o.g that $\sigma^2=1$, we have:
$$R(\theta,\bar{X})=\frac{1}{n}$$
By our hypothesis, there must exist an estimator δ' such that $R\left(\theta,\delta'\right)\leq1/n$ for all θ and $R\left(\theta',\delta'\right)<1/n$ for at least one $\theta'\in\Omega$. Because $R(\theta,\delta)$ is continuous in θ, there must exist $\varepsilon>0$ and an interval (θ_0,θ_1) containing θ' so that:
$$R\left(\theta,\delta'\right)<\frac{1}{n}-\varepsilon\quad\forall\theta\in(\theta_0,\theta_1).$$
Let r'_τ be the average risk of δ' with respect to the prior distribution $\mathcal{N}\left(0,\tau^2\right)$ on θ. (Note that this is the exact same prior we used to prove that $\bar{X}$ was the limit of a Bayes estimator) and hence minimax. We did this by letting $\tau\to\infty$, and therefore letting our prior tend to the improper prior $\pi(\theta)=1\forall\theta$.) Let r_τ be the average risk of a Bayes estimator δ_τ under the same prior.
Note that $\delta_\tau\neq\delta'$ because $R(\theta,\delta_\tau)\to\infty$ as $\theta\to\infty$ which is not consistent with $R\left(\theta,\delta'\right)\leq1/n$ for all $\theta\in\mathbb{R}$. So, $r_\tau<r'_\tau$, because the Bayes estimator is unique almost surely with respect to the marginal distribution of θ. We will look at the following ratio, which is selected to simplify our algebra later. This ratio, we will show, will become arbitrarily large, which we will use to form a contradiction with $r_\tau<r'_\tau$.
Using the form of the Bayes risk r_τ computed in a previous lecture (see TPE Example 5.1.14), we can write:
$$\frac{\frac{1}{n}-r'_\tau}{\frac{1}{n}-r_\tau}=\frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}\left[\frac{1}{n}-R\left(\theta,\delta'\right)\right]\exp\left(\frac{-\theta^2}{2\tau^2}\right)d\theta}{\frac{1}{n}-\frac{1}{n+\frac{1}{\tau^2}}}$$
Applying (11.1), we find:
$$\frac{\frac{1}{n}-r'_\tau}{\frac{1}{n}-r_\tau}\geq\frac{\frac{1}{\sqrt{2\pi}\tau}\int_{\theta_0}^{\theta_1}\varepsilon e^{\frac{-\theta^2}{2\tau^2}}d\theta}{\frac{1}{n\left(1+n\tau^2\right)}}$$
$$=\frac{n\left(1+n\tau^2\right)}{\tau\sqrt{2\pi}}\varepsilon\int_{\theta_0}^{\theta_1}e^{\frac{-\theta^2}{2\tau^2}}d\theta$$
As $\tau\to\infty$, the first expression, $n\left(1+n\tau^2\right)\varepsilon/(\tau\sqrt{2\pi})\to\infty$ and since the integrand converges monotonically to 1 , Lebesgue's monotone convergence theorem ensures that the integral approaches the positive quantity $\theta_1-\theta_0$. So, for sufficiently large τ, we must have
$$\frac{\frac{1}{n}-r'_\tau}{\frac{1}{n}-r_\tau}>1$$
This means that $r'_\tau<r_\tau$. However, this is a contradiction, because r_τ is the optimal average risk (since it is the Bayes risk). So our assumption that there was a dominating estimator was false, and in this case, $a\bar{X}+b=\bar{X}$ is admissible.

Mac-Zygmund SLLN for $0<p<1$: (Only change condition 2 as following)

2. $X_1,\ldots X_n\sim N(\mu_1,\sigma^2)$ $Y_1,\ldots Y_m\sim N(\mu_2,\sigma^2)$ σ^2 unknown

(a) $H_0:\mu_2\leq\mu_1$ $H_1:\mu_2>\mu_1$

(b) show UMP

$\gamma=\mu_2-\mu_1$

$H_0:\gamma\leq0$ $H_1:\gamma>0$

pooled sample variance $S^2=\frac{1}{m+n-2}\left(\sum_{i=1}^n(X_i-\bar{X})^2+\sum_{i=1}^m(Y_i-\bar{Y})^2\right)$

$\bar{X}\sim N(\mu_1,\frac{\sigma^2}{n})$ $\bar{Y}\sim N(\mu_2,\frac{\sigma^2}{m})$

$\bar{Y}-\bar{X}\sim N(\gamma,\sigma^2\frac{m+n}{nm})\Rightarrow\sqrt{\frac{nm}{m+n}}\left(\frac{\bar{Y}-\bar{X}-\gamma}{\sigma}\right)\sim N(0,1)$

plug in the estimate S for the unknown σ:

$T=\sqrt{\frac{nm}{m+n}}\left(\frac{\bar{Y}-\bar{X}-\gamma}{S}\right)\sim t_{m+n-2}$

$\phi(X,Y)=\begin{cases}1, & \text{if } T<-t_{m+n-2}(1-\alpha)\\ 0, & \text{if } T\geq-t_{m+n-2}(1-\alpha)\end{cases}$

where $t_\gamma(1-\alpha)=-t_\gamma(\alpha)$ is $(1-\alpha)$-quantile of t_γ-dist.

$X_i\sim U(0,\theta)$

$\mathbb{E}\left[\left(\frac{X_1}{X_{(n)}}\right)^k\right]=\mathbb{E}\left[\frac{X_1}{X_{(n)}}\right]^k\mathbb{E}\left[X_{(n)}\right]^{-k}=EX_1^k$

$\frac{X_{(n)}}{\theta}\sim\text{Beta}(n,1)$ $\frac{X_{(n)}}{\theta}\sim\text{Beta}(n,2)$

$EX_1^k=\theta^k\int_0^1\frac{n!}{(n-1)!0!}x^{n-1+k}dx$ $EX_1^k=\frac{\theta^{k+1}}{k+1}$

$EX^k=\frac{\theta^{k+1}}{(k+1)(k+2)}$

$Y_1=X_{(1)}$ $Y_k=X_k-X_{(k-1)}$ $i\geq2$

$\sum_{i=2}^n X_i-X_{(1)}=\sum_{i=2}^n Y_i$ $\perp\!\!\!\perp$ $Y_1=X_{(1)}$

$t_n=\frac{\sqrt{n}\bar{X}}{S_n}\xrightarrow{d}N(0,1)$

let $\sigma^2=E[X_1^2]$

By CLT, $\sqrt{n}\frac{\bar{X}_n}{\sigma}\xrightarrow{d}N(0,1)$

By SLLN, $S_n^2\to\sigma^2$ as

By Slutsky's theorem $\frac{\sqrt{n}(\bar{X}_n)}{\sigma}\cdot\frac{\sigma}{S_n}\xrightarrow{d}N(0,1)$

as $n\to\infty$ $P(-z_{1-\frac{\alpha}{2}}\leq\sqrt{n}\frac{\bar{X}_n-\mu}{S_n}\leq z_{1-\frac{\alpha}{2}})\to1-\alpha$

$X_1,X_2,\ldots X_n\sim f(x,\theta)=\frac{1}{2}e^{-|x-\theta|}$ $x\in\mathbb{R}$

Find MLE, Fisher Information

$L(X_i,\theta)=\prod_{i=1}^n f(X_i,\theta)=\prod_{i=1}^n\frac{1}{2}e^{-|X_i-\theta|}=\left(\frac{1}{2}\right)^n\exp\left(-\sum_{i=1}^n(X_i-\theta)\right)^2$

$\hat{\theta}=\arg\max_\theta L(X,\theta)=\arg\min_\theta\sum_{i=1}^n|X_i-\theta|$

$\hat{\theta}=\begin{cases}X_{(\frac{n+1}{2})}, & n\text{ odd}\\ \frac{X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}}{2}, & n\text{ even}\end{cases}$

$I_X(\theta)=E_\theta\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)^2=E_\theta(\text{sign}(\theta-X))^2=1$

$I_\theta=nI_X(\theta)=n$

[Question] countable add'y $\Rightarrow$ continuity + finite add'y

if $A_n\downarrow\varnothing$, that is $A_1\supset A_2\supset\cdots$ and $\bigcap_{i=1}^\infty A_i=\varnothing$, $A_i\in A_{i-1},\cdots$ take $B_i=A_i^c, i=1,\ldots$

then $B_n\uparrow\Omega$, that is $B_1\subset B_2\subset\cdots$ and $\bigcup_{i=1}^\infty B_i=\Omega$ construct $C_i=\begin{cases}B_1 & i=1\\ B_i\setminus B_{i-1} & i\geq2\end{cases}$

then $C_i, i=1,\ldots$ are disjoint and $\bigcup_{i=1}^\infty C_i=\bigcup_{i=1}^\infty B_i=\Omega$

$$\lim_{n\to\infty}P(B_n)=\lim_{n\to\infty}P\left(\bigcup_{i=1}^n B_i\right)=\lim_{n\to\infty}P\left(\bigcup_{i=1}^n C_i\right)=\lim_{n\to\infty}\sum_{i=1}^n P(a_i)$$
$$=\sum_{i=1}^\infty P(C_i)=P\left(\bigcup_{i=1}^\infty C_i\right)=P(\Omega)$$

then $\lim_{n\to\infty}P(A_n)=\lim_{n\to\infty}P(\Omega\setminus B_n)=\lim_{n\to\infty}[P(\Omega)-P(B_n)]$
$=P(\Omega)-\lim_{n\to\infty}P(B_n)=0$

$\Rightarrow$ continuity

continuity + finite add'y $\Rightarrow$ countable add'y.

Given a collection of set A, derived from Ω, satisfies finite additivity and continuity.

For $\forall$ disjoint $A_1,\cdots\in A$, take $A=\bigcup_{i=1}^\infty A_i$, $B_n=A\setminus\bigcup_{i=1}^n A_i$ then $B_1\supset B_2\supset\ldots\supset B_n\to\varnothing$ when $n\to\infty$ then according to the definition of continuity, $\lim_{n\to\infty}P(B_n)\to0$ by B_n and $\bigcup_{i=1}^n A_i$ disjoint then $P(A)=P\left(B_n\cup\left(\bigcup_{i=1}^n A_i\right)\right)=P(B_n)+P(\bigcup_{i=1}^n A_i)$

$$=P(B_n)+\sum_{i=1}^n P(A_i)\to0+\sum_{i=1}^\infty P(A_i)\text{ when }n\to\infty.$$

$\Rightarrow$ countable additivity

$|E\,S_n'|\leq\sum_{k=1}^n E|X_k|I\{|X_k|\leq k^{1/r}\}$
$=\sum_{k=1}^n E|X_k|I\{|X_k|\leq k^{1/(2r)}\}+\sum_{k=1}^n E|X_k|I\{k^{1/(2r)}<|X_k|\leq k^{1/r}\}$
$\leq\sum_{k=1}^n(k^{1/(2r)})^{1-r}E|X|^rI\{|X|\leq k^{1/(2r)}\}$
$+\sum_{k=1}^n(k^{1/r})^{1-r}E|X|^rI\{k^{1/(2r)}<|X|\leq k^{1/r}\}$
$\leq\sum_{k=1}^n k^{(1-r)/(2r)}E|X|^r+\sum_{k=1}^n k^{(1/r)-1}E|X|^rI\{|X|>k^{1/(2r)}\}$
$\leq Cn^{(1+r)/(2r)}E|X|^r+\sum_{k=1}^n k^{(1/r)-1}E|X|^rI\{|X|>k^{1/(2r)}\},$

where, in the last inequality, we glanced at Lemma A.3.1 without specifying the constant. Thus,

$$\frac{|E\,S_n'|}{n^{1/r}}\leq n^{(1/2)-(1/r)}\cdot E|X|^r+\frac{1}{n^{1/r}}\sum_{k=1}^n k^{(1/r)-1}E|X|^rI\{|X|>k^{1/(2r)}\},$$

which converges to 0 as $n\to\infty$, since the first term in the right-hand side converges to 0 (the exponent is negative), and the second term is a weighted average of terms tending to 0 (Lemma A.6.1).
If $f(x)$ can be written as $g_1(x_1)\cdots g_n(x_n)$ where the $g_m\geq0$ are measurable, then $X_1,X_2,\ldots,X_n$ are independent. Note that the g_m are not assumed to be probability densities.

Exercise 2.1.1. By Thm 2.1.8, it suffices to show that $P(X_1\leq z_1,\ldots,X_n\leq z_n)=\Pi_{i=1}^n P(X_i\leq z_i)$. Let $c_i=\int_{-\infty}^\infty g_i(x_i)dx_i$, for $i=1,\ldots,n$. Then, we have $c_1\cdots c_n=1$. Our claim is that
$$P(X_i\leq z_i)=\frac{1}{c_i}\int_{-\infty}^{z_i}g_i(x_i)dx_i.$$
WLOG, let $i=1$.
$$P(X_1\leq z_1)=P(-\infty<X_1\leq z_1,-\infty<X_2<\infty,\ldots,-\infty<X_n<\infty)$$
$$=(c_2\cdots c_n)\int_{-\infty}^{z_1}g_1(x_1)dx_1$$
$$=\frac{1}{c_1}\int_{-\infty}^{z_1}g_1(x_1)dx_1.$$
Now,
$$P(X_1\leq z_1,\ldots,X_n\leq z_n)=\left(\int_{-\infty}^{z_1}g_1(x_1)dx_1\right)\cdots\left(\int_{-\infty}^{z_n}g_n(x_n)dx_n\right)$$
$$=\left(\frac{1}{c_1}\int_{-\infty}^{z_1}g_1(x_1)dx_1\right)\cdots\left(\frac{1}{c_n}\int_{-\infty}^{z_n}g_n(x_n)dx_n\right)$$
$$=\prod_{i=1}^n P(X_i\leq z_i)$$

[Question] $P(X_1=x_1,\ldots,X_n=x_n)=\prod_{i=1}^n P(X_i=x_i)\Leftrightarrow X_1,\ldots,X_n$ independent.

Exercise 2.1.2. Let $\mathcal{A}_i=\{X_i^{-1}(x_i):x_i\in S_i\}\cup\{\emptyset,\Omega\}$. Let $I=\{n_1,\ldots,n_k\}\subset\{1,\ldots,n\}$ and $J=\{1,\ldots,n\}\setminus I=\{m_1,\ldots,m_{n-k}\}$. Then,
$$P(X_{n_1}=x_{n_1},\cdots,X_{n_k}=x_{n_k})=\sum_{x_{m_1}\in S_{m_1}}\cdots\sum_{x_{m_{n-k}}\in S_{m_{n-k}}}P(X_1=x_1,\cdots,X_n=x_n)$$
$$=\prod_{i\in I}P(X_i=x_i).$$
By Lemma 2.1.5, $\mathcal{A}_i$'s are independent. By Theorem 2.1.7, $\sigma(\mathcal{A}_i)$'s are independent. Since $X_i^{-1}(A)\in\sigma(\mathcal{A}_i)$ for all $A\in2^{S_i}$, $\sigma(X_i)\subset\sigma(\mathcal{A}_i)$. Therefore, $X_1,\ldots,X_n$ are independent.

b) $X_1,X_2,\ldots$ are iid with mean 0 and variance 1. Prove that
$$\lim_{n\to\infty}\frac{1}{n^{1/2}}\mathbb{E}\left[\sum_{i=1}^n|X_i|\right]=\sqrt{\frac{2}{\pi}}$$

(b) Let $Z\sim N(0,1)$. $E|Z|=2\int_0^\infty z\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz=\sqrt{\frac{2}{\pi}}$.

$\frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^n X_i\xrightarrow{d}Z\Rightarrow\frac{1}{n^{\frac{1}{2}}}E\left|\sum_{i=1}^n X_i\right|\to E|Z|$.

STAT5010

Generalized Inverse

decomposition:

1. full rank: $A_{m\times n} = B_{m\times r}\cdot C_{r\times n}$.
2. $\exists$ nonsingular P, Q s.t. $P\cdot A\cdot Q = (I_r)_{m\times n}$.
3. (symmetric) $\exists$ orthogonal P, s.t. $P^\top AP = D$.

idempotent: $A = A^2$; Orthogonal: $Q^\top Q = QQ^\top = I$

1. all idempotent matrices (except I) are singular.
2. if A is idempotent, then $\text{rank}(A) = \text{tr}(A)$.
3. if A is idempotent, then e.v. of A are 1 or 0.
4. if A is sym. with e.v. 1 and 0, then A is idempotent.

Pf: (2) $tr(A) = tr(BC) = tr(CB) = tr(I_{r\times r}) = r = r(A)$

inverse:

1. left inverse: $(r = n)$ $A^{-1}_{\text{left}} = (A^\top A)^{-1}A^\top$.
2. right inverse: $(r = m)$ $A^{-1}_{\text{right}} = A^\top(AA^\top)^{-1}$.
3. Moore-Penrose inverse:
   (a) AA^+ and A^+A are symmetric,
   (b) $A = AA^+A$ and $A^+ = A^+AA^+$.
4. generalized inverse: $AA^-A = A$.

Moore-Penrose inverse:

1. existence: $A^+ = C^\top(CC^\top)^{-1}(B^\top B)^{-1}B^\top$.
2. uniqueness: $AA_1^+ = (AA_2^+AA_1^+)^\top = AA_2^+$.
3. $\text{rank}(A) = \text{rank}(A^+)$.
4. $(A^\top)^+ = (A^+)^\top$. and if $A = A^\top$, then $A^+ = (A^+)^\top$.
5. $AA^+, A^+A, I_m - AA^+, I_n - A^+A$ are sym. and idempotent.
5. If A is nonsingular, $A^{-1} = A^+$.
6. If A is symmetric idempotent, $A^+ = A$.
7. If $r(A_{m\times n}) = m$, then $A^+ = A^\top(AA^\top)^{-1}, AA^+ = I$,
   If $r(A_{m\times n}) = n$, then $A^+ = (A^\top A)^{-1}A^\top, A^+A = I$.

generalized inverse: $AA^{-1}A = A$.

1. not unique.
2. Moore-Penrose inverse is also generalized inverse.
3. $\text{rank}(X^-) \geq r$, $\text{rank}(X^-X) = \text{rank}(XX^-) = r$.
4. $X^-X = I_n$ iff $\text{rank}(X) = n$. $XX^- = I_m$ iff $\text{rank}(X) = m$.
   (Hint: full-rank -> nonsingular (+ idempotent) -> identity.)
5. if G is G-inv. of $X^\top X$, then $G^\top$ is also G-inv of $X^\top X$.
6. ★ $XGX^\top X = X$, i.e. $GX^\top$ is G-inv. of X.
   take $K = XGX^\top$, then $KX = X, X^\top K = X^\top$
7. $K = XGX^\top$ is invariant w.r.t. the choice of G.
8. $K = XGX^\top$ is symmetric and idempotent.
9. $K = XGX^\top = XX^+$.

Proof:

(i) Since G is a generalized inverse of (X^TX), $(X^TX)G(X^TX) = X^TX$. Taking the transpose of both sides

$$\begin{aligned}\left[X^TX\right]^T &= \left[(X^TX)G(X^TX)\right]^T\\ &= (X^TX)^TG^T(X^TX)^T\end{aligned}$$

But $(X^TX)^T = X^T(X^T)^T = X^TX$, hence $(X^TX)G^T(X^TX) = (X^TX)$.

(ii) From (i) $(X^TX)G^T(X^TX) = (X^TX)$. Denote $(X^TX)G^T$ by B. Then

$$\begin{aligned}0 &= BX^TX - X^TX\\ &= (BX^TX - X^TX)(B^T - I)\\ &= BX^TXB^T - X^TXB^T - BX^TX - X^TX\\ &= (BX^T - X^T)(BX^T - X^T)^T\end{aligned}$$

Hence, $0 = BX^T - X^T$
$\Rightarrow \quad BX^T = X^T$
$\Rightarrow \quad X^TXG^TX^T = X^T$
Taking the transpose

$$\begin{aligned}X &= (X^TXG^TX^T)^T\\ &= XGX^TX\end{aligned}$$

Hence, GX^T is a generalized inverse for X.

(iii) Suppose F and G are generalized inverses for X^TX. Then, from (ii)

$$XGX^TX = X$$

and

$$XFX^TX = X$$

It follows that

$$\begin{aligned}0 &= X - X\\ &= (XGX^TX - XFX^TX)\\ &= (XGX^TX - XFX^TX)(G^TX^T - F^TX^T)\\ &= (XGX^T - XFX^T)X(G^TX^T - F^TX^T)\\ &= (XGX^T - XFX^T)(XG^TX^T - XF^TX^T)\\ &= (XGX^T - XFX^T)(XGX^T - XFX^T)^T\end{aligned}$$

Since the (i,i) diagonal element of the result of multiplying a matrix by its transpose is the sum of the squared entries in the i-th row of the matrix, the diagonal elements of the product are all zero only if all entries are zero in every row of the matrix. Consequently,

$$(XGX^T - XFX^T) = 0$$

(iv) For any generalized inverse G,

$$T = GX^TXG^T$$

is a symmetric generalized inverse. Then

$$XTX^T$$

is symmetric and from (iii),

$$XGX^T = XTX^T.$$

[Algorithm of Obtaining G-inv] Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank r, and $A_{11} \in \mathbb{R}^{r\times r}$. If A_{11} is invertible, then $G = \begin{pmatrix}A_{11}^{-1} & 0\\ 0 & 0\end{pmatrix} \in \mathbb{R}^{n\times m}$ is a G-inv of A.

Proof →

Note

$$\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}A_{11}^{-1} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix} = \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{21}A_{11}^{-1}A_{12}\end{bmatrix}.$$

By the definition of generalized inverse, it suffices to verify $A_{21}A_{11}^{-1}A_{12} = A_{22}$. Since the row rank of $[A_{11}, A_{12}]$ and A are both r, then there exists a $(n-r)\times r$-matrix P such that $P[A_{11}, A_{12}] = [A_{21}, A_{22}]$. And it can be found that $P = A_{21}A_{11}^{-1}$, and $PA_{12} = A_{22}$.

Distributions and Quadratic Forms

Random vector ($\mu = EY$)

1. $cov(Y) = E((Y-EY)(Y-EY)^T) = E(YY^T) - \mu\mu^T$.
2. $cov(AY) = A\,cov(Y)A^T$
3. $cov(AY, BY) = A\,cov(X, Y)B^T$.
4. $tr(ABC) = tr(BCA) = tr(CAB)$
   $Y^TAY = tr(Y^TAY) = tr(AYY^T)$
5. $E(Y^TAY) = tr(A\Sigma) + \mu^TA\mu$ ($cov(Y) = \Sigma$, $EY = \mu$)
   Proof: $E(Y^TAY) = E(tr(AYY^T)) = tr(E(AYY^T))$
   $= tr(A(\Sigma + \mu\mu^T)) = tr(A\Sigma) + \mu^TA\mu$.
6. $M_Y(t) = E(e^{t^TY})$
7. $Y_1, \ldots, Y_m$ indep. $\Rightarrow g_1(Y_1), \ldots, g_m(Y_m)$ indep.

MVN: (a) $f_Y(y) = |\Sigma|^{-\frac{1}{2}}(2\pi)^{-\frac{p}{2}}e^{-\frac{1}{2}((y-\mu)^T\Sigma^{-1}(y-\mu))}$
(b) $M_Y(t) = e^{t^T\mu + \frac{1}{2}t\Sigma}$.
(c) $Y \sim N(\mu, \Sigma) \Rightarrow BY + c \sim N(B\mu + c, B\Sigma B^T)$

Marginal, conditional, independence.

Let $Y = \begin{pmatrix}Y_1\\ Y_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right)$,
then (1) $Y_1 \sim N(\mu_1, \Sigma_{11})$.
(2) $Y_1 \mid Y_2 = y_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
(3) $Y_1 \perp\!\!\!\perp Y_2$ iff $\Sigma_{12} = 0$.

distribution-χ^2: $X \sim N(0, I_n)$, $X^TX \sim \chi^2_{(n)}$.

1. if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_n)$, then $\mathbf{x}'\mathbf{x} \sim \chi^2_{(n,\lambda)}$ with $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}$.
2. m.g.f. $(1-2t)^{-\frac{n}{2}}\exp\{-\lambda(1 - \frac{1}{1-2t})\}$.
3. $w = \dfrac{u_1/p_1}{u_2/p_2} \sim F(p_1, p_2, \lambda)$, where $u_1 \sim \chi^2_{(p_1,\lambda)}$, $u_1 \sim \chi^2_{p_2}$.
4. $t = \dfrac{z}{\sqrt{\frac{u}{n}}} \sim t_{(n,\mu)}$, if $z \sim N(\mu, 1)$, $u \sim \chi^2_{(n)}$, $z \perp\!\!\!\perp u$.

positive definite matrix.

1. eigenvalues of a p.d. matrix is positive.
2. Symetric A is p.d. iff $A = P^TP$ where P is nonsingular.
3. Symmetric $A_{n\times n}$, $r(A) = r$. $A = LL^T$ where $L_{n\times r}$ has rank r.

quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

1. m.g.f.: $|\mathbf{I} - 2t\mathbf{A}\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}\boldsymbol{\mu}'[\mathbf{I} - (\mathbf{I} - 2t\mathbf{A}\boldsymbol{\Sigma})^{-1}]\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}$.
2. If $\mathbf{A}$ & $\mathbf{V}$ sym., $\mathbf{V}$ p.d., then e.v. of $\mathbf{AV}$ 0, 1 -> $\mathbf{AV}$ idem.

Proof: Recall that that a symmetric matrix is positive definite if and only if it can be written as $P^\top P$ for a nonsingular P. Then, $V = P^\top P$ for some nonsingular matrix P. If $|AV - \lambda I| = 0$ has roots 0 and 1, then

$$\begin{aligned}&|P||AV - \lambda I||P^{-1}| = 0 && \text{has roots } 0 \,\&\, 1\\ \Rightarrow\ &|PAVP^{-1} - \lambda PIP^{-1}| = 0 && \text{has roots } 0 \,\&\, 1\\ \Rightarrow\ &|PAP^\top PP^{-1} - \lambda I| = 0 && \text{has roots } 0 \,\&\, 1\\ \Rightarrow\ &|PAP^\top - \lambda I| = 0 && \text{has roots } 0 \,\&\, 1. \qquad (1)\end{aligned}$$

Thus, $PAP^\top$ has eigenvalues 0 and 1. But $PAP^\top$ is symmetric (because A is symmetric). Hence, $PAP^\top$ is idempotent. That is

$$PAP^\top PAP^\top = PAP^\top \Rightarrow PAVAP^\top = PAP^\top.$$

As P is nonsingular,

$$P^{-1}PAVAP^\top P = P^{-1}PAP^\top P$$

which implies

$$AVAV = AV.$$

3. If $\mathbf{A}$ sym. idem. with rank r, then $\mathbf{A}$ has r e.v.-1 and rest -0.

THEOREM 1. *Let $x_{p\times 1} \sim N(\mu, \Sigma)$ and let A be symmetric. Then, $q = x^\top Ax$ follows $\chi^2_{(r,\lambda)}$ with r being the rank of A and $\lambda = \frac{\mu^\top A\mu}{2}$ if and only of $A\Sigma$ is idempotent.*

Corollaries:

1. If $x \sim N(0, I)$, then $x^\top Ax$ is χ^2_r if and only if A is idempotent of rank r.
2. If $x \sim N(0, V)$, then $x^\top Ax$ is χ^2_r if and only if AV is idempotent of rank r.
3. If $x \sim N(\mu, \sigma^2 I)$, then $\frac{x^\top x}{\sigma^2}$ is $\chi^2_{(n, \frac{1}{2\sigma^2}\mu^\top\mu)}$.
4. If $x \sim N(\mu, I)$, the $x^\top Ax$ is $\chi^2_{(r, \frac{1}{2}\mu^\top A\mu)}$ if and only if A is idempotent of rank r.

Properties of normal random vector:

(1) $E(x^\top Ax) = tr(A\Sigma) + \mu^\top A\mu$;

(2) $Cov(x, x^\top Ax) = 2\Sigma A\mu$;

(3) $Var(x^\top Ax) = 2tr[(AV)^2] + 4\mu^\top A\Sigma A\mu$.

THEOREM 2. *Let $x \sim N_n(\mu, \Sigma)$. Then $x^\top Ax$ and Bx are distributed independently if and only if $B\Sigma A = 0$.*

Proof: *(1. Sufficiency)* To prove $B\Sigma A = 0$ implies independence. Since A is symmetric, one can write $A = LL^\top$ for some L of full column rank. If $B\Sigma A = 0$, then $B\Sigma LL^\top = 0$. Next, as L has full column rank, $(L^\top L)^{-1}$ exists. Thus,

$$B\Sigma LL^\top = 0 \Rightarrow B\Sigma LL^\top L(L^\top L)^{-1} = 0 \Rightarrow B\Sigma L = 0.$$

Therefore,

$$Cov(Bx, x^\top L) = BCov(x)L = B\Sigma L = 0.$$

Hence, since x is a vector of normally distributed random variables, Bx and $x^\top L$ are independently normally distributed which implies Bx and $x^\top Ax = x^\top LL^\top x$ are distributed independently.

(2. Necessity) To prove independence of $x^\top Ax$ and Bx implies $B\Sigma A = 0$. By independence, $Cov(Bx, x^\top Ax) = 0$. On the other hand, $Cov(Bx, x^\top Ax) = 2B\Sigma A\mu$. Hence,

$$2B\Sigma A\mu = 0.$$

Since this is true for all μ, $B\Sigma A = 0$ and the proof is complete.

THEOREM 3. *Let $x \sim N(\mu, \Sigma)$. Then $x^\top Ax$ and $x^\top Bx$ are distributed independently if and only if $B\Sigma A = 0$(or equivalently, $A\Sigma B = 0$).*

Proof: *(1. Sufficiency)* That $A\Sigma B = 0$ implies independence. By Lemma 1, we can write $A = LL^\top$ and $B = MM^\top$ where each L and M have full column rank. Hence, if $A\Sigma B = 0$, then $LL^\top\Sigma MM^\top = 0$. Since $(L^\top L)^{-1}$ and $(M^\top M)^{-1}$ exist,

$$(L^\top L)^{-1}L^\top(LL^\top\Sigma MM^\top)M(M^\top M)^{-1} = 0,$$

which implies

$$L^\top\Sigma M = 0.$$

Therefore, $Cov(L^\top x, x^\top M) = L^\top\Sigma M = 0$. Thus, as x is a vector of normally distributed variables, $L^\top x$ and $x^\top M$ are distributed independently. Consequently, $x^\top Ax = x^\top LL^\top x$ and $x^\top Bx = x^\top MM^\top x$ are distributed independently.

matrix identity:

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}^- = \begin{bmatrix}A^- + A^-B(D - CA^-B)^-CA^- & -A^-B(D - CA^-B)^-\\ -(D - CA^-B)^-CA^- & (D - CA^-B)^-\end{bmatrix}$$
$$= \begin{bmatrix}(A - BD^-C)^- & -(A - BD^-C)^-BD^-\\ -D^-C(A - BD^-C)^- & D^- + D^-C(A - BD^-C)^-BD^-\end{bmatrix}$$

Woodbury matrix identity:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}.$$

THEOREM 4. *Let $x \sim N(\mu, V)$ and let A_i be $n\times n$ symmetric matrix of rank k_i, for $i = 1, 2, \cdots, p$. Denote $A = \sum_{i=1}^p A_i$, which is symmetric with rank k. Then,*

$$x^\top A_ix \sim \chi^2_{(k_i, \frac{1}{2}\mu^\top A_i\mu)},$$

$x^\top A_ix$ are pairwise independent and $x^\top Ax$ is $\chi^2_{(k, \frac{1}{2}\mu^\top A\mu)}$, if and only if
(I): any 2 of
(a) A_iV idempotent, for all i;
(b) $A_iVA_j = 0$ for all $i < j$;
(c) AV is idempotent.
are true;
or (II): (c) is true and (d) $k = \sum_{i=1}^p k_i$;
or(III): (c) is true and (e) $A_1V, \cdots, A_{p-1}V$ are indempotent and A_pV is non-negative definite.

Derivatives: $X = (X_1, \ldots, X_p)^T$.
① $\frac{\partial f}{\partial X} = \left(\frac{\partial f}{\partial X_1} \cdots \frac{\partial f}{\partial X_p}\right)^T$. ② $\frac{\partial L^TX}{\partial X} = L$ ③ $\frac{\partial X^TL}{\partial X} = L$
④ $\frac{\partial X^TAX}{\partial X} = AX + A^TX$ (if $A = A^T$ then $= 2AX$)

Linear Regression for the full-rank model

ordinary least squares estimation

1. loss function: $L(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.
2. normal equation: $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$.
3. LSE: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
4. hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
   - symmetric and idempotent.
   - projection, $\mathbf{HX} = \mathbf{X}$, $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
5. residuals: $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.
   $\mathbb{E}[\text{SSE}] = \mathbb{E}[\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}] = (n-p)\sigma^2$. $\frac{SSE}{\sigma^2} \sim \chi^2_{n-r(x)}$

The hat matrix H is symmetric idempotent;
Proof: Note that $(X^\top X)$ is symmetric and $(X^\top X)^{-1}$ is also symmetric. Then,

$$H = X(X^\top X)^{-1}X^\top = H^\top, \qquad HH = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = H.$$

2. $X^\top\hat{\varepsilon} = 0$; (This holds because of $X^\top H = X^\top$, $HX = X$ and $X^\top(I - H) = 0$, $(I - H)X = 0$.)

Proof: Since

$$X^\top H = X^\top X(X^\top X)^{-1}X^\top = X^\top,$$
$$HX = X(X^\top X)^{-1}X^\top X = X,$$

then,

$$X^\top(I - H) = X^\top - X^\top H = X^\top - X^\top = 0,$$
$$(I - H)X = X - HX = X - X = 0.$$

Clearly,

$$X^\top\hat{\varepsilon} = X^\top(I - H)Y = 0.$$

3. $\hat{Y}^\top\hat{\varepsilon} = 0$;

Proof: Write

$$(HY)^\top(I - H)Y = Y^\top H^\top(I - H)Y = Y^\top H(I - H)Y$$
$$= Y^\top 0Y = 0.$$

4. $I - H$ is symmetric idempotent;
5. $E(\hat{\beta}) = \beta_0$ (unbiased estimate);

Proof:

$$E(\hat{\beta}) = E((X^\top X)^{-1}X^\top Y) = (X^\top X)^{-1}X^\top E(Y) = (X^\top X)^{-1}X^\top X\beta_0 = \beta_0.$$

6. $Cov(\hat{\beta}) = (X^\top X)^{-1}\sigma^2$;

Proof:

$$\begin{aligned}Cov(\hat{\beta}) &= Cov((X^\top X)^{-1}X^\top Y) = (X^\top X)^{-1}X^\top Cov(Y)X(X^\top X)^{-1}\\ &= \sigma^2(X^\top X)^{-1}X^\top IX(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.\end{aligned}$$

7. $tr(I_n - H) = n - p$; ★ $\frac{1}{p}\sum Var(\hat\beta_j) = \frac{1}{p}\sigma^2 tr((X^TX)^{-1}) = \frac{1}{p}\sigma^2\sum\lambda_i^{-2}$ (SVD)

Proof: Note that

$$tr(H) = tr(X(X^\top X)^{-1}X^\top) = tr(X^\top X(X^\top X)^{-1}) = tr(I_p) = p.$$

Then,

$$tr(I_n - H) = tr(I_n) - tr(H) = n - p.$$

8. $\hat{\varepsilon}^\top\hat{\varepsilon} = tr(YY^\top(I - H))$;

Proof: We can easily show that

$$\begin{aligned}\hat{\varepsilon}^\top\hat{\varepsilon} &= Y^\top(I - H)^\top(I - H)Y = Y^\top(I - H)Y\\ &= tr(Y^\top(I - H)Y) = tr(YY^\top(I - H)).\end{aligned}$$

9. $E(\mathbf{Y}\mathbf{Y}^\top) = \sigma^2 I + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}^\top$;
10. $\hat{\varepsilon}^\top\hat{\varepsilon}/(n-p)$ is an unbiased estimate of σ^2, that is

$$E\left(\frac{\hat{\varepsilon}^\top\hat{\varepsilon}}{n-p}\right) = \sigma^2.$$

Proof: Write

$$\begin{aligned}E(\hat{\varepsilon}^\top\hat{\varepsilon}) &= E(Y^\top(I - H)(I - H)Y) = E(Y^\top(I - H)Y)\\ &= tr((I - H)\Sigma) + \beta^\top X^\top(I - H)X\beta\\ &= \sigma^2 tr(I - H) = \sigma^2(n - p).\end{aligned}$$

weighted least squares estimation

1. assumption: $\boldsymbol{\Sigma}$ is known.
2. loss function: $S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.
3. WLSE: $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$

best linear unbiased estimator:

1. $\begin{cases}\text{minimizes } \text{Var}(\boldsymbol{\lambda}'\mathbf{Y}) = \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}\\ \text{s.t. } \mathbb{E}[\boldsymbol{\lambda}'\mathbf{Y}] = \mathbf{t}'\boldsymbol{\beta} \Rightarrow \mathbf{X}'\boldsymbol{\lambda} = \mathbf{t}\end{cases}$
2. using $2\boldsymbol{\theta}$ as Lagrange multiplier, and then minimize
   $W(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{\theta}'(\mathbf{X}'\boldsymbol{\lambda} - \mathbf{t})$.

$$\frac{\partial W(\boldsymbol{\lambda}, \boldsymbol{\theta})}{\partial\boldsymbol{\lambda}} = 2\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{X}\boldsymbol{\theta} = 0,$$
$$\frac{\partial W(\boldsymbol{\lambda}, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = 2\boldsymbol{X}^\top\boldsymbol{\lambda} - 2\boldsymbol{t} = 0.$$

Solving the above two equations for $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, we have

$$\boldsymbol{\lambda}^\top = \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}.$$

Therefore, the BLUE of $\boldsymbol{t}^\top\boldsymbol{\beta}$ is

$$\boldsymbol{\lambda}^\top\boldsymbol{Y} = \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

$= t^T\hat\beta_{ws}$

with variance

$$\begin{aligned}Var(\boldsymbol{\lambda}^\top\boldsymbol{Y}) &= \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma})(\boldsymbol{\Sigma})^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{t}\\ &= \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{t}.\end{aligned}$$

Remark 5. In a special case that $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, the BLUE of $\boldsymbol{t}^\top\boldsymbol{\beta}$ is

$$\boldsymbol{t}^\top(\boldsymbol{X}^\top(\boldsymbol{I}\sigma^2)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}\sigma^2)^{-1}\boldsymbol{Y} = \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}, \quad = t\hat\beta_{ols}$$

with variance

$$\boldsymbol{t}^\top(\boldsymbol{X}^\top(\boldsymbol{I}\sigma^2)^{-1}\boldsymbol{X})^{-1}\boldsymbol{t} = \sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}.$$

Remark 6. By letting $\boldsymbol{t}^\top$ be, in turn, each row of $\boldsymbol{I}_k$, we can easily obtain the BLUE of $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$, which is precisely the weighted least square estimate or generalized least square estimate.

Remark 7. When $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, the BLUE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$.

MLE

$$L(\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}e^{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}.$$
$$\frac{\partial\log L}{\partial\beta} = \frac{1}{\sigma^2}(\boldsymbol{X}^\top\boldsymbol{Y} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}),$$
$$\frac{\partial\log L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$
$$\text{MLE}(\beta) = \tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}.$$
$$\text{MLE}(\sigma^2) = \tilde{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \frac{1}{n}[\boldsymbol{Y}^\top\boldsymbol{Y} - \tilde{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{Y}].$$

1. The MLE of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$ follows $N(\beta, (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sigma^2)$.
2. The sum of squares of the deviations of the observed $Y_i's$ from their estimated expected values is usually known as the *residual error sum of squares* or *sum of squares due to error*, denoted as SSE. It is given by
   $$SSE = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$
   Then, $\tilde{\sigma}^2 = \frac{SSE}{n}$ is the MLE. But we have shown that $\hat{\sigma}^2 = \frac{SSE}{n - r(\boldsymbol{X})}$ is an unbiased estimator of σ^2 with $r(\boldsymbol{X}) = $ rank of $\boldsymbol{X}$, implying $\tilde{\sigma}^2$ is biased for finite n. However, it is consistent or asymptotically unbiased as $n \to \infty$.
3. The MLE $\hat{\boldsymbol{\beta}}$ and SSE are independent. To show this, recall that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$ and $SSE = \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ where $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Since
   $$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\sigma^2\boldsymbol{I})(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0},$$
   by Theorem 2 in section 2.4, $\hat{\boldsymbol{\beta}}$ and SSE are independent and thus $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.
4. Distribution of $\hat{\sigma}^2$. Consider $\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}\sigma^2)$.
   Since
   $$\left(\frac{\boldsymbol{I} - \boldsymbol{H}}{\sigma^2}\right)(\boldsymbol{I}\sigma^2) = \boldsymbol{I} - \boldsymbol{H}$$
   which is idempotent, then
   $$\frac{SSE}{\sigma^2} \sim \chi^2_{r(\frac{I-H}{\sigma^2}), \frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta})^\top(\frac{I-H}{\sigma^2})(\boldsymbol{X}\boldsymbol{\beta})}.$$
   However,
   $$(\boldsymbol{X}\boldsymbol{\beta})^\top\left(\frac{\boldsymbol{I} - \boldsymbol{H}}{\sigma^2}\right)(\boldsymbol{X}\boldsymbol{\beta}) = 0.$$
   Therefore, the noncentrality parameter is 0. In addition,
   $$r\left(\frac{\boldsymbol{I} - \boldsymbol{H}}{\sigma^2}\right) = r(\boldsymbol{I} - \boldsymbol{H}) = n - r(\boldsymbol{X}).$$
   Consequently,
   $$\frac{SSE}{\sigma^2} \sim \chi^2_{n - r(\boldsymbol{X})}, \qquad \text{or}$$
   $$\frac{[n - r(\boldsymbol{X})]\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n - r(\boldsymbol{X})}.$$

[Lemma] If
$$\boldsymbol{M} = \begin{bmatrix}\boldsymbol{X}^\top\\ \boldsymbol{Z}^\top\end{bmatrix}[\boldsymbol{X} \quad \boldsymbol{Z}] = \begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{Z}\\ \boldsymbol{Z}^\top\boldsymbol{X} & \boldsymbol{Z}^\top\boldsymbol{Z}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{B}^\top & \boldsymbol{D}\end{bmatrix},$$
$$\boldsymbol{W} = (\boldsymbol{D} - \boldsymbol{B}^\top\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} = [\boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{Z}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}]^{-1},$$

[Deviations from Means] $Y_{n\times 1} = (1 \quad X_1)\begin{pmatrix}\beta_0\\ b^T\end{pmatrix} + \varepsilon_{n\times 1}$.
Let $S = X_1^TX_1 - n\bar{x}\bar{x}^T = Z^TZ$ and $Z = X_1 - 1\bar{x}^T$.
$\hat\beta_0 = \bar{y} - \bar{x}^T\hat{b} = \frac{1}{n}1^TY - \bar{x}^T\hat{b}$.
$\hat{b} = S^{-1}(X_1^TY - n\bar{y}\bar{x}) = (Z^TZ)^{-1}Z^TY$.
$var(\hat\beta_0) = \frac{\sigma^2}{n} + \bar{x}^Tvar(\hat{b})\bar{x}^T$.
$var(\hat{b}) = (Z^TZ)^{-1}\sigma^2$. $cov(\hat\beta_0, b^T) = -\bar{x}^Tvar(\hat{b})$.

ANOVA

$SST = Y^TY$, $\frac{SST}{\sigma^2} \sim \chi^2\left(n, \frac{\frac{1}{2}\beta^TX^TX\beta}{\sigma^2}\right)$
$SSR = Y^THY = \hat\beta^TX^TY$, $\frac{SSR}{\sigma^2} \sim \chi^2\left(r(x), \frac{\frac{1}{2}\beta^TX^TX\beta}{\sigma^2}\right)$
$SSE = Y^T(I-H)Y = (Y-\hat{Y})^T(Y-\hat{Y})$. $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-r(x), 0)}$
$F(R) = \frac{SSR}{r(x)} \Big/ \frac{SSE}{n-r(x)} \sim F\left(r(x), n-r(x), \frac{\beta^TX^TX\beta}{2\sigma^2}\right)$
under $H_0: \beta = 0$. $F(R) = F(r, n-r; 0)$

Hypothesis Testing $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$. (s×(k+1), s×1)

1. $(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}) \sim \mathcal{N}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}, \sigma^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})$.
2. $\mathbf{Q} = (\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})' \left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} (\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}) \sim \sigma^2\chi^2_{(s,\lambda)}$.
3. $\mathbf{Q}$ and SSE are independent.
   Hint: $\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{Y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}\right]$.
4. $F(H) = \dfrac{\mathbf{Q}/s}{\mathrm{SSE}/(n - \mathrm{r}(\mathbf{X}))} \overset{H_0}{\sim} F_{s,n-\mathrm{r}(\mathbf{X})}$.

Estimation under constraint (Null hypothesis):

1. use 2θ as Lagrange multiplier
2. minimize $W(\boldsymbol{\beta},\boldsymbol{\theta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\theta}'(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})$,
3. get $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})$.

$SSE_{H_0}(\tilde{\beta}) = SSE(\hat{\beta}) + Q \geq SSE(\hat{\beta}) = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$

· (Properties of $\beta^0 = GX^\top y$) (1) $E(\beta^0) = GX^\top X\beta$. (2) $\mathrm{Var}(\beta^0) = GX^\top XG^\top\sigma^2$. (3) $\hat{y} = X\beta^0 = XGX^\top y$. (4) $E(\hat{y}) = XGX^\top X\beta = X\beta$. (5) $\mathrm{SSE} = y^\top(I - XGX^\top)y$. (6) $SSR = y^\top XGX^\top y = (\beta^0)^\top X^\top y$. (7) $\hat{\sigma}^2 = \frac{SSE}{n - r(X)}$ is an unbiased estimator of σ^2. (8) $SSR_m = y^\top(XGX^\top - \frac{11^\top}{n})y$.

· (Distributional Properties) (1) $\beta^0 \sim N(GX^\top X\beta, GX^\top XG^\top\sigma^2)$. The covariance matrix is singular. (2) $\frac{SSE}{\sigma^2} \sim \chi^2_{n-r(X)}$, $\frac{SSR}{\sigma^2} \sim \chi^2(r(X), \frac{1}{\sigma^2}\beta^\top X^\top X\beta)$, $\frac{MSR}{MSE} \sim F(r(X), n - r(X), \frac{1}{2\sigma^2}\beta^\top X^\top X\beta)$.

· (Identifiability) Formally, the parameter β is identifiable if $f(\beta_1) = f(\beta_2)$ implies that $\beta_1 = \beta_2$ for any β_1 and β_2. More generally, the vector-valued function $g(\beta)$ is identifiable if $f(\beta_1) = f(\beta_2)$ implies that $g(\beta_1) = g(\beta_2)$.

[Proposition] In a linear model for which X is of full rank, β is identifiable.

[Proposition] A function $g(\beta)$ is identifiable if and only if $g(\beta) = (h \circ f)(\beta)$ for some function h. $\beta_1 \neq \beta_2 \Rightarrow \mu_1 \neq \mu_2$.

· (Estimable functions) If a vector t exists such that $t^\top E(y) = q^\top\beta$, then $q^\top\beta$ is said to be estimable. Linear combinations of estimable functions are estimable.

· $E(t^\top y) = t^\top X\beta = q^\top\beta \Rightarrow t^\top X = q^\top$ for some t. This is equivalent to saying that q is in the row space of X.

Theorem 1. (Gauss-Markov Theorem) The best linear unbiased estimator of the estimable function $q^\top\beta$ is $q^\top\beta^0$.

Proof. We prove this theorem from the following three aspects:

(i) $q^\top\beta^0 = q^\top GX^\top y \iff$ linear function of y_i.

(ii)

$$\begin{aligned} E(q^\top\beta^0) &= q^\top E(\beta^0) \\ &= q^\top GX^\top E(y) \\ &= q^\top GX^\top X\beta \\ &= t^\top XGX^\top X\beta = t^\top X\beta = q^\top\beta. \end{aligned}$$

$\Rightarrow$ unbiased estimator.

(iii) Minimum variance. (Optimality)

$$\begin{aligned} var(q^\top\beta^0) &= q^\top var(\beta^0)q \\ &= q^\top GX^\top XG^\top q\sigma^2 \\ &= t^\top XGX^\top XG^\top X^\top t\sigma^2 \\ &= t^\top XGX^\top t\sigma^2 \\ &= q^\top Gq\sigma^2. \end{aligned}$$

Suppose $k^\top y$ is another linear unbiased estimator of $q^\top\beta$ different from $q^\top\beta^0$.

$$\begin{aligned} &\Rightarrow E(k^\top y) = q^\top\beta \\ &\Rightarrow k^\top X = q^\top \\ &\Rightarrow cov(q^\top\beta^0, k^\top y) \\ &= cov(q^\top GX^\top y, k^\top y) \\ &= q^\top GX^\top(I\sigma^2)k \\ &= q^\top Gq\sigma^2. \end{aligned}$$

$$\begin{aligned} & var(q^\top\beta^0 - k^\top y) \\ =& var(q^\top\beta^0) + var(k^\top y) - 2cov(q^\top\beta^0, k^\top y) \\ =& var(k^\top y) + q^\top Gq\sigma^2 - 2q^\top Gq\sigma^2. \\ \Rightarrow\ & var(q^\top\beta^0 - k^\top y) \\ =& var(k^\top y) - var(q^\top\beta^0) \geq 0 \\ \Rightarrow\ & var(k^\top y) \geq var(q^\top\beta^0) \\ \Rightarrow\ & q^\top\beta^0 \text{ is B.L.U.E. of } q^\top\beta. \end{aligned}$$

Remark: Under normality assumption, $q^\top\beta^0 \sim N(q^\top\beta, q^\top Gq\sigma^2)$.

[Theorem] (Test of Estimability) The linear function $q^\top\beta$ is estimable if and only if $q^\top H = q^\top$ where $H = GX^\top X$.

· (Testable hypothesis) A testable hypothesis is $H_0 : K^\top\beta = m$ such that $K^\top = T^\top X$ for some matrix $T^\top$ of order $r \times n$. The matrix $K^\top$ of size $r \times p$ is always of full-row rank.

· (Hypothesis testing) $K^\top\beta^0 - m \sim N(K^\top\beta - m, K^\top GK\sigma^2)$, $Q = (K^\top\beta^0 - m)^\top(K^\top GK)^{-1}(K^\top\beta^0 - m)$. Then $\frac{Q}{\sigma^2} \sim \chi^2(s, (K^\top\beta - m)^\top(K^\top GK)^{-1}(K^\top\beta - m)/2\sigma^2)$, $F(H) = \frac{Q/s}{SSE/(n - r(X))} \sim F(s, n - r(X), (K^\top\beta - m)^\top(K^\top GK)^{-1}(K^\top\beta - m)/2\sigma^2)$. $\supset 0$ under H_0.

$\beta^0_H = \beta^0 - GK(K^\top GK)^{-1}(K^\top\beta^0 - m)$

$SSE_H = SSE + Q$.

Bias-Var: $\mathrm{PE}(z_0) = E_{Y|Z=z_0}[(Y - \hat{f}(Z))^2 \mid Z = z_0] = \sigma^2_\varepsilon + \mathrm{Bias}^2(\hat{f}(z_0)) + \mathrm{Var}(\hat{f}(z_0)) = \sigma^2_\varepsilon + \mathrm{MSE}(\hat{f}(z_0))$.

Ridge Regression

· (Ridge regression) Minimize $(Y - Z\beta)^\top(Y - Z\beta)$ s.t. $\sum_{j=1}^p \beta_j^2 \leqslant t$. $\hat{\beta}^{\text{ridge}} = \arg\min_\beta \|Y - Z\beta\|_2^2 + \lambda\|\beta\|_2^2$. Its solution may have smaller average PE than $\hat{\beta}^{ls}$. Note that $\mathrm{PRSS}(\beta)_{\ell_2}$ is convex, and hence has a unique solution. $\hat{\beta}^{\text{ridge}}_\lambda = (Z^\top Z + \lambda\mathrm{I}_p)^{-1}Z^\top Y$.

· By convention, (1) Z is assumed to be standardized (mean $= 0$, variance $= 1$) and (2) Y is assumed to be centered.

· ($\hat{\beta}^{\text{ridge}}_\lambda$ is biased). Let $R = Z^\top Z$ and Z has full column rank. $\hat{\beta}^{\text{ridge}}_\lambda = (Z^\top Z + \lambda\mathrm{I}_p)^{-1}Z^\top Y = (R + \lambda\mathrm{I}_p)^{-1}R(R^{-1}Z^\top Y) = [R(\mathrm{I}_p + \lambda R^{-1})]^{-1}R\hat{\beta}^{ls} = (\mathrm{I}_p + \lambda R^{-1})^{-1}\hat{\beta}^{ls}$. Thus, $E(\hat{\beta}^{\text{ridge}}_\lambda) = E[(\mathrm{I}_p + \lambda R^{-1})^{-1}\hat{\beta}^{ls}] = (\mathrm{I}_p + \lambda R^{-1})^{-1}\beta \neq \beta$ $(\forall\beta \neq 0, \lambda \neq 0)$ Note that 1 is not an eigenvalue of the matrix $(\mathrm{I}_p + \lambda R^{-1})^{-1}$ as $\lambda \neq 0$ and R is positive definite.

· (SVD) By the singular value decomposition, $Z = UDV^\top$ where U is an $n \times p$ orthogonal matrix, D is $p \times p$ diagonal matrix and $V^\top$ is a $p \times p$ orthogonal matrix. $\hat{\beta}^{\text{ridge}}_\lambda = V\,\mathrm{diag}(\frac{d_j}{d_j^2 + \lambda})U^\top Y$, by using $Z^\top Z = (UDV^\top)^\top(UDV^\top) = VD^\top U^\top UDV^\top = VD^2V^\top$.

· (Orthonormal Z) Z is orthonormal, then $Z^\top Z = I_p$. Let $\hat{\beta}^{ls}$ denote the LS solution, then $\hat{\beta}^{\text{ridge}}_\lambda = \frac{1}{1+\lambda}\hat{\beta}^{ls}$ and $\hat{\beta}^{\text{ridge}}_j = \frac{\hat{\beta}^{ols}_j}{1+\lambda}$. The optimal choice of λ minimizing the expected prediction error is $\lambda^* = \frac{p\sigma^2}{\sum_{j=1}^p \beta_j^2}$ where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the true coefficient vector.

· Variance of the ridge regression estimate is $\mathrm{Var}(\hat{\beta}^{\text{ridge}}) = \sigma^2 W_\lambda(Z^\top Z)^{-1}W_\lambda = \sigma^2(Z^\top Z + \lambda I_p)^{-1}Z^\top Z(Z^\top Z + \lambda I_p)^{-1}$ where $W_\lambda = (Z^\top Z + \lambda I_p)^{-1}Z^\top Z$. Recall that $\hat{\beta}^{\text{ridge}} = W_\lambda\hat{\beta}^{ls}$

· The bias of the ridge regression estimate is $\mathrm{bias}(\hat{\beta}^{\text{ridge}}) = -\lambda W_\lambda(Z^\top Z)^{-1}\beta = -\lambda(Z^\top Z + \lambda I_p)^{-1}\beta$.

Ridge → OLS: $Z_\lambda = \begin{pmatrix} Z \\ \sqrt{\lambda}\mathrm{I}_p \end{pmatrix}$, $Y_\lambda = \begin{pmatrix} Y \\ 0 \end{pmatrix}$

Ridge: $E[\hat{\beta}_\lambda] = V\,\mathrm{diag}(\frac{d_i^2}{d_i^2+\lambda})V^T\beta$

$\mathrm{cov}[\hat{\beta}_\lambda] = \sigma^2 V\,\mathrm{diag}(\frac{d_i^2}{(d_i^2+\lambda)^2})V^T$

LASSO

· (LASSO) Minimize $(Y - Z\beta)^\top(Y - Z\beta)$, s.t. $\sum_{j=1}^p |\beta_j| \leqslant t$. It is often convenient to rewrite the LASSO problem in the so-called Lagrangian form : $\mathrm{minimize}_\beta(Y - Z\beta)^\top(Y - Z\beta) + \lambda\|\beta\|_1$, where $\lambda \geq 0$ is the so-called tuning parameter. The resulting lasso estimate is denoted by $\hat{\beta}^{\text{lasso}}_\lambda$.

· There is usually a factor $1/(2n)$ or $1/n$ appearing in front of $(Y - Z\beta)^\top(Y - Z\beta)$. This kind of standardization makes λ values comparable for different sample sizes (useful for cross-validation).

· We typically center the response and standardize the predictors so that each column is centered ($\frac{1}{n}\sum_{i=1}^n Z_{ij} = 0$) and has unit variance ($\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 = 1$).

· (KKT condition) The necessary and sufficient conditions for a solution to LASSO take the form

$$-\langle z_j, Y - Z\beta\rangle + \lambda s_j = 0, \quad j = 1, \ldots, p$$

Here each s_j is an unknown quantity equal to $\mathrm{sign}(\beta_j)$ if $\beta_j \neq 0$ and some value lying in $[-1, 1]$ otherwise.

· (Orthonormal Z) Consider an orthogonal design case with $Z^\top Z = \mathrm{I}$. The LASSO method is equivalent to : solve β_j's componentwisely by solving : $\min_{\beta_j}(\beta_j - \hat{\beta}^{ls}_j)^2 + \lambda|\beta_j|$. The solution to the above problem is

$$\hat{\beta}^{\text{lasso}}_j = \mathrm{sgn}(\hat{\beta}^{ls}_j)(|\hat{\beta}^{ls}_j| - \frac{\lambda}{2})_+ = \begin{cases} \hat{\beta}^{ls}_j - \frac{\lambda}{2} & \text{if } \hat{\beta}^{ls}_j > \frac{\lambda}{2} \\ 0 & \text{if } |\hat{\beta}^{ls}_j| \leqslant \frac{\lambda}{2} \\ \hat{\beta}^{ls}_j + \frac{\lambda}{2} & \text{if } \hat{\beta}^{ls}_j < -\frac{\lambda}{2} \end{cases}$$

[Question] $\exists\lambda > 0$, such that $MSE(\hat{\beta}^{\text{ridge}}(\lambda)) \leq MSE(\hat{\beta})$

$MSE(\hat{\beta}(\lambda)) = tr(Var(\hat{\beta}(\lambda))) + (E(\hat{\beta}(\lambda)) - \beta)^T(E(\hat{\beta}(\lambda)) - \beta)$ (C_2, C_1)

$C_1 = \lambda^2\sum_{j=1}^p \frac{\gamma_j^2}{(d_j^2+\lambda)^2}$, $C_2 = \sigma^2\sum_{j=1}^p \frac{d_j^2}{(d_j^2+\lambda)^2}$, where

$\gamma = V^T\beta = (\gamma_1, \ldots, \gamma_p)^T$

$MSE(\beta(\lambda)) = \lambda^2\sum_{j=1}^p \frac{\gamma_j^2}{(d_j^2+\lambda)^2} + \sigma^2\sum_{j=1}^p \frac{d_j^2}{(d_j^2+\lambda)^2}$

$\frac{\partial MSE(\lambda)}{\partial\lambda} = 2\lambda\sum_{j=1}^p \frac{\gamma_j^2 d_j^2}{(d_j^2+\lambda)^3} - 2\sigma^2\sum_{j=1}^p \frac{d_j^2}{(d_j^2+\lambda)^3} < 0$ for a λ close to 0.

Hence $\exists\lambda^* \in (0, \delta)$, s.t. $MSE(\beta(\lambda^*)) \leq MSE(\beta(0))$.

Under $X^TX = I$, then $\lambda^*\sum_{j=1}^p \frac{\hat{\beta}_j^2}{(1+\lambda^*)^3} = \hat{\sigma}^2\sum_{j=1}^p \frac{1}{(1+\lambda^*)^3}$, $\lambda^* = \frac{p\hat{\sigma}^2}{\|\hat{\beta}\|^2}$.

[Question] (1) Find $\hat{\theta}^{ols}$ for $L(\theta) = (Y - \theta)^T(Y - \theta) + \lambda(1^T\theta - 2\pi)$, $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 2\pi$

$\frac{\partial L}{\partial\theta} = -2(Y - \theta) - \lambda 1 = 0$; $\frac{\partial L}{\partial\lambda} = 1^T\theta - 2\pi = 0$.

$\hat{\theta} = Y + \frac{1}{4}(2\pi 1 - 11^T Y)$, $\hat{\theta}_i = Y_i + \frac{1}{4}(2\pi - \sum_{j=1}^4 Y_j)$

(2) Find unbiased $\hat{\sigma}^2$.

Since $\varepsilon = Y - \theta \sim N(0, \sigma^2 I)$, $1^T\varepsilon = 1^T(Y - \theta) = 1^TY - 2\pi \sim N(0, 4\sigma^2 I)$.

Since $4\sigma^2 = Var(1^T\varepsilon) = E(1^TY - 2\pi)^2$, $\frac{1}{4}(1^TY - 2\pi)^2$ is unbiased for σ^2.

(3) Test $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$.

Let $K^T = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$, test $K^T\theta = 0$. ($(I - \frac{1}{a}J)$ is idempotent)

$\hat{\theta} \sim N(\theta, \sigma^2(I - \frac{1}{4}J))$, $K^T\hat{\theta} \sim N(0, 2\sigma^2 I)$

$Q = \frac{(K^T\hat{\theta})^T(K^T\hat{\theta})}{2\sigma^2}$, $s = 2$; $SSE = (Y - \hat{\theta})^T(Y - \hat{\theta})$, $n - r = 1$, $\frac{Q/s}{SSE/(n-r)} \sim F(2, 1)$.

2. Consider the model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, i = 1,2,3,4, j = 1,2,3,4,$$

where ε_{ij} are independently distributed as $N(0, \sigma^2)$.

(1) Let $\beta = (\mu, \tau_1, \tau_2, \tau_3, \tau_4)'$. Find a set of 4 linearly independent estimable functions of β.

(2) Derive a test to test the null hypothesis $H_0 : \tau_1 - \tau_2 = \tau_3 - \tau_4$.

(3) Is $\tau_1 + 2\tau_2$ estimable? Why?

(1) $X = \begin{pmatrix} 1_{4\times1} & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}_{16\times5}$ $X^TX = \begin{pmatrix} 16 & 4 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 4 & 0 \\ 4 & 0 & 0 & 0 & 4 \end{pmatrix}$

$G = (X^TX)^- = \mathrm{diag}(0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

$q_1^T = (1\ 1\ 0\ 0\ 0)$, $q_2^T = (1\ 0\ 1\ 0\ 0)$, $q_3^T = (1\ 0\ 0\ 1\ 0)$, $q_4^T = (1\ 0\ 0\ 0\ 1)$

$q_i^T\beta$ is estimable by choosing $t_i^TX = q_i^T$.

(b) $K^T = (0\ 1\ -1\ -1\ 1)$, test $K^T\beta = 0$.

Let $Q = (K^T\hat{\beta})^T(K^T(X^TX)^-K)^{-1}(K^T\hat{\beta})$, where $\hat{\beta} = GX^Ty$,

under H_0, $Q \sim \sigma^2\chi^2_{(1)}$, $SSE \sim \sigma^2\chi_{n-r(X)}$, $r(X) = 4$.

$F(H) = \frac{Q/s}{SSE/(n - r(X))} \sim F(1, 12)$

(c) $H = GX^TX = \begin{pmatrix} 0 & & & & \\ 1 & 1 & & & \\ 1 & & 1 & & \\ 1 & & & 1 & \\ 1 & & & & 1 \end{pmatrix}$. $q^TH = (3\ 1\ 2\ 0\ 0) \neq q^T$ Hence $\tau_1 + 2\tau_2$ not estimable.

(a) Let $A_{m\times m}$ and $B_{n\times n}$ be two nonsingular matrices. Further, assume that the matrices U and V are $m \times n$ and $n \times m$ respectively. Prove that 左右都直接乘

$$(A + UBV)^{-1} = (A^{-1} - A^{-1}UB(B + BVA^{-1}UB)^{-1}BVA^{-1}.)$$

$(A + UBV)$

(b) Consider a regression model,

$$Y = X\beta + \epsilon,$$

where X, $n \times p$, is full column rank. $Y = (Y_1, \ldots, Y_n)'$. Further, assume that $Var(\epsilon) = \sigma^2 I$. Let e_i be the ith residual and h_i be the ith diagonal element of the hat matrix. Let $\hat{\beta}$ and $\hat{\beta}_{(i)}$ be the least squares estimate of β with and without the ith case included in the data respectively.

i. Show that

$$(X'_{(i)}X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1 - h_i},$$

where $X_{(i)}$ denotes the regression matrix with the ith row (x_i') deleted. (Hint: $X'X = X'_{(i)}X_{(i)} + x_ix_i'$) $X' = X_{(i)} + (0, \ldots, 0, x_i, 0, \ldots, 0)$

ii. Prove that

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1}x_ie_i}{1 - h_i}.$$

(a) $(A + UBV)(A^{-1} - A^{-1}UB(B + BVA^{-1}UB)^{-1}BVA^{-1})$
$= I + UBVA^{-1} - U((B + BVA^{-1}UB)(B + BVA^{-1}UB)^{-1}BVA^{-1}) = I + UBVA^{-1} - UBVA^{-1} = I$.

(b) i. $(X_{(i)}^TX_{(i)})^{-1} = (X^TX - x_ix_i^T)^{-1}$. Let $A = X^TX$, $U = x_i$, $B = -I$, $V = x_i^T \Rightarrow (X_{(i)}^TX_{(i)})^{-1} = (X^TX)^{-1} + (1 - x_i^T(X^TX)^{-1}x_i)^{-1}(X^TX)^{-1}x_ix_i^T(X^TX)^{-1} = (X^TX)^{-1} + (1 - h_i)^{-1}(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}$

ii. It suffices to prove $\left(I + \frac{(X^TX)^{-1}x_ix_i^T}{1 - h_i}\right)\hat{\beta} = \left((X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1 - h_i}\right)(X^Ty - y_ix_i) + \frac{(X^TX)^{-1}x_iy_i}{1 - h_i}$

$RHS = (X^TX)^{-1}X^Ty - (X^TX)^{-1}x_iy_i + (X^TX)^{-1}x_ix_i^T(X^TX)^{-1}X^Ty/(1 - h_i) - (X^TX)^{-1}x_ix_i^T(X^TX)^{-1}x_iy_i/(1 - h_i) + (X^TX)^{-1}x_iy_i/(1 - h_i)$
$= \hat{\beta} - (X^TX)^{-1}x_iy_i + (X^TX)^{-1}x_ix_i^T\hat{\beta}/(1 - h_i) - (X^TX)^{-1}x_iy_ih_i/(1 - h_i) + (X^TX)^{-1}x_iy_i/(1 - h_i)$
$= \hat{\beta} + \frac{(X^TX)^{-1}x_ix_i^T\hat{\beta}}{1 - h_i} + \underbrace{(-1 - \frac{h_i}{1 - h_i} + \frac{1}{1 - h_i})}_{=0}(X^TX)^{-1}x_iy_i = LHS$

3. There are two groups of data (Y_1, X_1), (Y_2, X_2) with

$$Y_1 = X_1\beta + \varepsilon_1,$$
$$Y_2 = X_2\beta + \varepsilon_2,$$

X_1 and X_2 is not necessarily full-rank. And suppose that $\lambda^T\beta$ is estimable.

(1) T_1 and T_2 are BLUE of $\lambda^T\beta$ for data (Y_1, X_1) and (Y_2, X_2), respectively. Give T_1 and T_2 and calculate $Var(T_1)$ and $Var(T_2)$

(2) Let $T(\alpha) = \alpha T_1 + (1 - \alpha)T_2$. Find α to minimize $Var(T(\alpha))$

(3) Let $Y = (Y_1^T, Y_2^T)^T$, $X = (X_1^T, X_2^T)^T$, give the BLUE T_3 of $\lambda^T\beta$ for data (Y, X). And calculate $Var(T_3)$.

(4) Explain $Var(T_3) \leq Var(T(\alpha))$ with equality when $r(X_1) = 1$ or $r(X_2) = 1$

(1) $\lambda^T\beta^0 = \lambda^TGX^Ty$ is the BLUE of $\lambda^T\beta$. $G = (X^TX)^-$.
$var(T_i) = var(\lambda^T\beta_i^0) = \lambda^TG_i\lambda\sigma^2$, $G_i = (X_i^TX_i)^-$, $i = 1, 2$.

(2). $var(T(\alpha)) = \alpha^2 var(T_1) + (1 - \alpha)^2 var(T_2)$
$= (var(T_1) + var(T_2))\alpha^2 - 2var(T_2)\alpha + var(T_2)$
$\alpha^* = var T_2/(var T_1 + var T_2) = \lambda^TG_2\lambda/\lambda^T(G_1 + G_2)\lambda$

(3) $T_3 = \lambda^TGX^Ty$, $G = (X_1^TX_1 + X_2^TX_2)^-$, $var T_3 = \lambda^TG\lambda\sigma^2$.

(4) $var(T(\alpha^*)) = \frac{var T_1\, var T_2}{var(T_1) + var(T_2)}$

To prove $\lambda^T(X_1^TX_1 + X_2^TX_2)^-\lambda \leq \frac{\lambda^T(X_1^TX_1)^-\lambda\,\lambda^T(X_2^TX_2)^-\lambda}{\lambda^T((X_1^TX_1)^- + (X_2^TX_2)^-)\lambda}$, use

(5) A and B are symmetric and nonnegtive matrix and a is in the vector space of A and B, that's to say, there exist x and y s.t. $a = Ax = By = (A + B)z$, then

$$a^TA^-aa^TB^-a \geq a^T(A + B)^-a(a^TA^-a + a^TB^-a)$$

with equality if $r(A) = 1$ or $r(B) = 1$.

$a^TA^-a = x^TPP^Tx$; $a^TB^-a = y^TQQ^Ty$
NTS: $x^TAx\,y^TBy \geq z^T(A+B)z(x^TAx + y^TBy)$
$(z^TAx)^2 \leq (z^TAz)(x^TAx)$; $(z^TBy)^2 \leq (z^TBz)(y^TBy)$
Since $z^TAx = z^TBy = z^T(A+B)z$,
$z^TAz + z^TBz \leq (z^T(A+B)z)^2(\frac{1}{x^TAx} + \frac{1}{y^TBy})$,
$x^TAx\,y^TBy \geq z^T(A+B)z(x^TAx + y^TBy)$,

If $r(A) = 1$, then $A = uu^T$, where u is a vector. Then

$$(z^Tuu^Tx)^2 = (z^Tuu^Tz)(x^Tuu^Tx) \quad (z^TAx)^2 = (z^TAz)(x^TAx)$$

$$\begin{aligned} By &= Ax = uu^Tx = ku \\ Bz &= Ax - Az = (k - l)u \end{aligned}$$

where $k = u^Tx$, $l = u^Tz$. Thus, we have $Bz = QQ^Tz = \frac{k-l}{k}QQ^Ty$, then $Q^Tz = \frac{k-l}{k}Q^Ty$ s.t.

$$(z^TQQ^Ty)^2 = (z^TQQ^Ty)(x^TQQ^Ty). \quad (z^TBy)^2 = (z^TAz)(x^TAx)$$

[Question]

i. Consider the following linear model:

$$\begin{aligned} Y_{ijt} &= \gamma_i + \tau_j + \epsilon_{ijt}, \quad (1) \\ \epsilon_{ijt} &\sim N(0, \sigma^2_E), \gamma_i \sim N(0, \sigma^2_\gamma), i = 1, 2; j = 1, 2; t = 1, 2; \end{aligned}$$

where all random variables on the right hand side of the model are mutually independent. Write the model as $Y = Z\gamma + X\tau + \epsilon$, where

$$Y = [Y_{111}, Y_{112}, Y_{121}, Y_{122}, Y_{211}, Y_{212}, Y_{221}, Y_{222}], \gamma = [\gamma_1, \gamma_2], \tau = [\tau_1, \tau_2]$$

and find Z, X. Next, find the variance-covariance matrix of $Z\gamma + \epsilon$.

Distribution of Y:

Given independence and normality of γ_i and ϵ_{ijt}:

$$Y \sim N(X\tau, \sigma^2_\gamma ZZ^T + \sigma^2_E I_8).$$

Best linear unbiased estimator (BLUE) for τ:

Using generalized least squares (GLS), the BLUE for τ is found as:

$$\hat{\tau} = (X^T(\sigma^2_\gamma ZZ^T + \sigma^2_E I_8)^{-1}X)^{-1}X^T(\sigma^2_\gamma ZZ^T + \sigma^2_E I_8)^{-1}Y.$$

[Question] $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, $\hat{\beta}$ is OLS of $Y = X\beta + \varepsilon$. $\hat{\beta}^*$ is OLS of $Y_1 = X_1\beta^* + \varepsilon^*$; $\hat{Y} = X\hat{\beta}$; $\hat{Y}^* = X\hat{\beta}^*$.
$Y - \hat{Y} = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ $Y - \hat{Y}^* = \begin{pmatrix} e_1^* \\ e_2^* \end{pmatrix}$.

(a) Show $\hat{\beta} - \hat{\beta}^* = (X_1^TX_1)^{-1}X_2^Te_2$ $\Leftarrow\Rightarrow$

$\hat{\beta} = \hat{\beta}^* + (X_1^TX_1)^{-1}X_2^Te_2 = (X_1^TX_1)^{-1}(X_1^Ty_1 + X_2^Ty_2 - X_2^TX_2\hat{\beta})$

$(I + (X_1^TX_1)^{-1}(X_2^TX_2))\hat{\beta} = (X_1^TX_1)^{-1}X^Ty$.

$(X_1^TX_1)(I + (X_1^TX_1)^{-1}(X_2^TX_2))(X^TX)^{-1}X^Ty = X^Ty$. ✓

(b). $e_2 = y_2 - \hat{y}_2 = y_2 - \hat{y}_2^* + \hat{y}_2^* - \hat{y}_2 = e_2^* + X_2(\hat{\beta}^* - \hat{\beta})$
$\hat{\beta} - \hat{\beta}^* = (X_1^TX_1)^{-1}X_2^T(e_2^* - X_2(\hat{\beta} - \hat{\beta}^*))$
$\hat{\beta} - \hat{\beta}^* = (I + (X_1^TX_1)^{-1}X_2^TX_2)^{-1}(X_1^TX_1)^{-1}X_2^Te_2^*$.

(c) show that $\|\hat{\beta}^{\text{ridge}}\|$ increases as the tuning parameter $\lambda \to 0$.

By the singular value decomposition, $Z = UDV^\top$, where U is an orthogonal matrix, D is a diagonal matrix. Then, $\hat{\beta}^{\text{ridge}}_\lambda$ can be expressed as

$$\hat{\beta}^{\text{ridge}}_\lambda = (Z^\top Z + \lambda I_p)^{-1}Z^\top Y = V\,\mathrm{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)U^\top Y.$$

Thus, we have

$$\|\hat{\beta}^{\text{ridge}}\|_2^2 = Y^\top Z[(Z^\top Z + \lambda I_p)^{-1}]^2Z^\top Y = Y^\top U\,\mathrm{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)^2U^\top Y.$$

Then, for any $\lambda_1 > \lambda_2$,

$$\left(\frac{d_j}{d_j^2 + \lambda_1}\right)^2 - \left(\frac{d_j}{d_j^2 + \lambda_2}\right)^2 < 0,$$

Thus,

$$\mathrm{diag}\left[\left(\frac{d_j}{d_j^2 + \lambda_1}\right)^2 - \left(\frac{d_j}{d_j^2 + \lambda_2}\right)^2\right]$$

is negative definite. Therefore, we obtain $\|\hat{\beta}^{\text{ridge}}\|_2$ increases as $\lambda \to 0$.

Show that the smallest value of λ such that the regression coefficients estimated by the lasso are all equal to zero is given by

[Question] $$\lambda_{\max} = \max_j\left|\frac{1}{N}\langle x_j, y\rangle\right|.$$

By KKT condition, $-\langle x_j, Y - X\beta\rangle + \lambda s_j = 0$, $j = 1, \ldots, p$. Then,

$$s_j = \mathrm{sgn}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0,$$
$$\text{or } s_j \in [-1, 1] \text{ if } \hat{\beta}_j = 0.$$

For any $0 < \epsilon < \lambda_{\max}$, by the definition of $\lambda_{\max}$, we know $\hat{\beta}^{\text{Lasso}}_{\lambda_{\max}-\epsilon} \neq 0$ but $\lim_{\epsilon\to0}\hat{\beta}^{\text{Lasso}}_{\lambda_{\max}-\epsilon} = 0$. Hence, we have

$$(\lambda_{\max} - \epsilon)s = \frac{1}{N}X^\top(Y - X^\top\hat{\beta}^{\text{Lasso}}_{\lambda_{\max}-\epsilon}) \to \lambda_{\max} - \epsilon = \frac{1}{N}\left\|X^\top(Y - X^\top\hat{\beta}^{\text{Lasso}}_{\lambda_{\max}-\epsilon})\right\|_\infty$$

Where the last equality holds since $\|s\|_\infty = 1$ by the fact $\hat{\beta}^{\text{Lasso}}_{\lambda_{\max}-\epsilon} \neq 0$. Thus, let $\epsilon \to 0$, we have

$$\lambda_{\max} = \frac{1}{N}\|X^\top Y\|_\infty = \max_j\left|\frac{1}{N}\langle x_j, y\rangle\right|$$

1. (Underfitted Model)

True Model: $Y = Xb + Z\gamma + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$.

Mis-specified Model: $Y = Xb + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$.

Let the least squares estimate of b be $\hat{b} = (X^\top X)^{-1}X^\top Y$,

(a) Prove that $\hat{b}$ is a biased estimator of b and find the bias.
(b) Find the variance of $\hat{b}$.
(c) Prove that S^2 is a biased estimator of σ^2 where
$$S^2 = \frac{Y^\top(I - X(X^\top X)^{-1}X^\top)Y}{n - r(X)}.$$
(d) Show that S^2 overestimate σ^2.
(e) Find $E(\hat{\varepsilon})$ and $Var(\hat{\varepsilon})$.

1. (a) $E(\hat{b}) = (X^\top X)^{-1}X^\top(Xb + Z\gamma)$, this is a biased estimator and the bias is $(X^\top X)^{-1}X^\top Z\gamma$.
(b) $Var(\hat{b}) = (X^\top X)^{-1}X^\top Var(Y)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}$.
(c)
$$E(S^2) = E\left[\frac{Y^\top(I - X(X^\top X)^{-1}X^\top)Y}{n - r(X)}\right] = \sigma^2 + \frac{(Z\gamma)^\top(I - X(X^\top X)^{-1}X^\top)(Z\gamma)}{n - r(X)}.$$
(d) $E(S^2) = \sigma^2 + \frac{(Z\gamma)^\top(I - H)(Z\gamma)}{n - r(X)}$ is overestimated.
(e) $E(\hat{\varepsilon}) = (I - X(X^\top X)^{-1}X^\top)(Z\gamma)$.
$Var(\hat{\varepsilon}) = \sigma^2(I - X(X^\top X)^{-1}X^\top)$

2. (Over-fitted model)

True Model: $Y = X_1b_1 + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$.

Mis-specified Model: $Y = Xb + \varepsilon$, $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$.

(a) What is $E(\hat{b})$.
(b) Evaluate $E(S^2)$.
(c) Evaluate $Var(\hat{b})$.

2. (a) $E(\hat{b}) = (X^\top X)^{-1}X^\top X_1b_1$.
(b) $E(S^2) = \sigma^2 + \frac{(X_1b_1)^\top(I - X(X^\top X)^{-1}X^\top)(X_1b_1)}{n - r(X)}$. Since $(I - H)X = (I - H)(X_1, X_2) = 0$, then $(I - H)X_1 = 0$. Therefore $E(S^2) = \sigma^2$.
(c) $Var(\hat{b}) = \sigma^2(X^\top X)^{-1}$.

8. Consider the linear regression model

$$y = X_{n\times p}\beta_{p\times1} + \varepsilon.$$

where $\varepsilon \sim N(0, \sigma^2 I)$. Let $\hat{\beta}$ be the least squares estimate of β. Define $\tilde{\beta} = c\hat{\beta}$ where $c \leq 1$. The mean squared error(MSE) of $\tilde{\beta}$ is

$$MSE(\tilde{\beta}) = E(\tilde{\beta} - \beta)^\top(\tilde{\beta} - \beta).$$

(a) Prove that $MSE(\tilde{\beta}) = c^2\sigma^2 tr(X^\top X)^{-1} + (c - 1)^2\beta^\top\beta$.
(b) Let c^* be the value of c such that $MSE(\tilde{\beta})$ is a minimum. Find c^*.
(c) Let $p = 5, \sigma^2 = 1, \beta^\top = (1, 2, 3, 4, 5)$ and the eigenvalues of $X^\top X$ be $1, 2, 3, 4, 5$. Evaluate c^*.

8. (a)

$$\begin{aligned} MSE(\tilde{\beta}) &= E(c(X^\top X)^{-1}X^\top Y - \beta)^\top(c(X^\top X)^{-1}X^\top Y - \beta) \\ &= tr(c^2\sigma^2(X^\top X)^{-1}) + (c - 1)^2\beta^\top\beta \\ &= c^2\sigma^2 tr((X^\top X)^{-1}) + (c - 1)^2\beta^\top\beta. \end{aligned}$$

(b) $MSE(\tilde{\beta}) = tr((X^\top X)^{-1}) + (c - 1)^2\beta^\top\beta$. Then minimizer $c^* = \frac{\beta^\top\beta}{\sigma^2 tr((X^\top X)^{-1}) + \beta^\top\beta}$.
(c) $c^* = \frac{3300}{3437}$.

(a) **Testing linear hypothesis**

Recall that in parametric linear model with normal error assumption, $Y \sim N(X\beta, \sigma^2 I)$. Then

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y, \qquad \hat{\beta} \sim N(\beta, (X^\top X)^{-1}\sigma^2).$$

$\Rightarrow K^\top\hat{\beta} - m \sim N[K^\top\beta - m, K^\top(X^\top X)^{-1}K\sigma^2]$. Next, since

$$\frac{(K^\top(X^\top X)^{-1}K)^{-1}}{\sigma^2}(K^\top(X^\top X)^{-1}K)\sigma^2 = I \quad \text{(symmetric and idempotent)},$$

therefore, $(K^\top(X^\top X)^{-1}K)^{-1}$ is symmetric (why?). Thus, we define

$$Q = (K^\top\hat{\beta} - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\hat{\beta} - m).$$

Then,

$$\frac{Q}{\sigma^2} \sim \chi^2_{s, \frac{1}{2\sigma^2}(K^\top\beta - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\beta - m)}.$$

(unknown; under H_0: $K^\top\beta = m$ → 0.)

Next, we intend to show that Q and SSE are independent.

$$\begin{aligned} Q &= (K^\top\hat{\beta} - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\hat{\beta} - m) \\ &= (K^\top(X^\top X)^{-1}X^\top Y - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top(X^\top X)^{-1}X^\top Y - m). \end{aligned}$$

But

$$K^\top(X^\top X)^{-1}X^\top Y - m = K^\top(X^\top X)^{-1}X^\top(Y - XK(K^\top K)^{-1}m).$$

(a shift of Y. $=: b$.)

Therefore,

$$Q = (Y - XK(K^\top K)^{-1}m)^\top X(X^\top X)^{-1}K(K^\top(X^\top X)^{-1}K)^{-1}K^\top(X^\top X)^{-1}X^\top(Y - XK(K^\top K)^{-1}m).$$

($d^\top$, B, d)

Note that if we let

$$d = Y - XK(K^\top K)^{-1}m,$$

$$\begin{aligned} Q &= d^\top[X(X^\top X)^{-1}K(K^\top(X^\top X)^{-1}K)^{-1}K^\top(X^\top X)^{-1}X^\top]d \\ &\equiv d^\top Bd, \end{aligned}$$

and

$$d \sim N(X\beta - XK(K^\top K)^{-1}m, \sigma^2 I).$$

On the other hand,

$$\begin{aligned} SSE &= Y^\top(I - X(X^\top X)^{-1}X^\top)Y \\ &= (Y - XK(K^\top K)^{-1}m)^\top(I - X(X^\top X)^{-1}X^\top)(Y - XK(K^\top K)^{-1}m) \\ &= d^\top[I - X(X^\top X)^{-1}X^\top]d \\ &\equiv d^\top Ad \end{aligned}$$

this is because

$$X^\top(I - X(X^\top X)^{-1}X^\top) = 0,$$
$$(I - X(X^\top X)^{-1}X^\top)X = 0.$$

Since

$$A\Sigma B = (I - X(X^\top X)^{-1}X^\top)\sigma^2 IX(X^\top X)^{-1}K[K^\top(X^\top X)^{-1}K]^{-1}K^\top(X^\top X)^{-1}X^\top = 0,$$

We have shown Q and SSE are independent. Finally, the test statistics

$$\begin{aligned} F(H) &= \frac{Q/s}{SSE/[N - r(X)]} \\ &= \frac{Q}{s\hat{\sigma}^2} \\ &\sim F'_{s, N-r(X), \frac{1}{2\sigma^2}(K^\top\beta - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\beta - m)} \end{aligned}$$

where $\hat{\sigma}^2 = \frac{SSE}{N - r(X)}$ is the unbiased estimator of σ^2.
Under H_0: $K^\top\beta = m$,

$$F(H) \sim F'_{(s, N-r(X))}.$$

3.6.1 Quantile.

Any real-valued random variable Y may be characterized by its (right-continuous) distribution function

$$F(y) = P(Y \le y),$$

whereas for any $0<\tau<1$,

$$F^{-1}(\tau) = \inf\{y: F(y) \ge \tau\}$$

is called the τth quantile of Y. The median $F^{-1}(1/2)$ plays the central role.

3.6.2 Linear quantile regression.

Given a covariate vector $X \in R^{p+1}$, the τth conditional quantile function of $Y \in R$ is modeled as

$$Q_{Y|X}(\tau) = X^\top \beta_\tau \quad (1)$$

for certain specific $\tau \in (0,1)$ of interest, and β_τ is $(p+1)$−vector usually including an intercept. For model (1) with complete data, a classical estimate of β_τ, denoted as $\hat{\beta}_\tau^c$, is obtained by minimizing

$$R(\beta_\tau) \equiv \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta_\tau) \quad (2)$$

over β_τ, where (y_i, x_i) are iid copies of (Y, X), $i=1,\dots,n$,

$$\rho_\tau(u) = u(\tau - I(u<0))$$

directional derivative of R in direction w is given by

$$\nabla R(\beta_\tau, w) \equiv \frac{d}{dt}R(\beta_\tau + tw)|_{t=0} = \frac{d}{dt}\sum_{i=1}^n [y_i - x_i^\top(\beta_\tau + tw)]\{\tau - I[y_i - x_i^\top(\beta_\tau + tw) < 0]\}|_{t=0} = -\sum_{i=1}^n \psi_\tau(y_i - x_i^\top\beta_\tau, -x_i^\top w)x_i^\top w,$$

where

$$\psi_\tau(u,v) = \begin{cases} \tau - I(u<0) & \text{if } u \ne 0; \\ \tau - I(v<0) & \text{if } u = 0. \end{cases}$$

3.6.3 Asymptotic properties of quantile regression.

The conditional distribution function of Y_is will be written as

$$P(Y_i < y|x_i) = F_{Y_i}(y|x_i) = F_i(y)$$

and so

$$Q_{Y_i}(\tau|x_i) = F_{Y_i}^{-1}(\tau|x_i) \equiv \xi_i(\tau).$$

We will employ the following regularity conditions to explore the asymptotic behavior of the estimator $\hat{\beta}_n(\tau)$, the minimizer of

$$\sum_{i=1}^n \rho_\tau(y_i - x_i^\top\beta_\tau).$$

- **Condition (A1).** The distribution functions $\{F_i\}$ are absolutely continuous, with continuous densities $f_i(\xi)$ uniformly bounded away from 0 and ∞ at the points $\xi_i(\tau), i = 1, 2, \dots$.
- **Condition (A2).** There exists positive definite matrices D_0 and $D_1(\tau)$ such at
  (i) $\lim_{n\to\infty} n^{-1}\sum x_i x_i^\top = D_0$;
  (ii) $\lim_{n\to\infty} n^{-1}\sum f_i(\xi_i(\tau))x_i x_i^\top = D_1(\tau)$;
  (iii) $\max_{i=1,\dots,n}\|x_i\|/\sqrt{n} \to 0$.

Conditions $A2(i)$ and $A2(iii)$ are familiar throughout the literature on M-estimators for regression models; some variant of them is necessary to ensure that a Lindeberg condition is satisfied. Condition $A2(ii)$ is really a matter of notational convenience and could be deduced from Condition $A2(i)$ and a slightly strengthened version of Condition $A1$.

Theorem 4.1 *Under Conditions A1 and A2,*

$$\sqrt{n}\{\hat{\beta}_n(\tau) - \beta_0(\tau)\} \to N(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1})$$

in distribution as $n\to\infty$. In particular, in the iid error model (that is, $Y_i = x_i^\top\beta + u_i$, $i = 1,\dots,n$ with iid errors u_i have a common distribution function F),

$$\sqrt{n}\{\hat{\beta}_n(\tau) - \beta_0(\tau)\} \to N(0, \omega^2 D_0^{-1}),$$

in distribution as $n\to\infty$, where $\omega^2 = \tau(1-\tau)/f_i^2(\xi_i(\tau))$.

Remark: The asymptotic properties above were established under fixed design, which can be extended to handle random design without further difficulties.

3.6.4 The least absolute deviation.

In the iid error model (random design) $Y_i = x_i^\top\beta + u_i$, $i=1,\dots,n$, where u_i have a common distribution function F, the LAD is a special case of quantile regression when $\tau = 0.5$. Its estimate $\hat{\beta}_{LAD}$ is defined to be the minimizer of

$$\min_\beta \sum_{i=1}^n |Y_i - X_i^\top\beta|.$$

Hence, the following theorem can be established in a straightforward fashion under regularity conditions.

Theorem 4.2 *Under Conditions A1 and A2,*

$$\sqrt{n}\{\hat{\beta}_{LAD} - \beta_0\} \to N(0, \omega^2 V^{-1})$$

in distribution as $n\to\infty$, where $\omega^2 = 1/4f^2(0)$ and $V = E(XX^\top)$.

Motivation of SEM:
① Group highly correlated observed variables into latents
② access interrelationships among latents through a regression

Model interpretation with latent ω_k
(1) $\Lambda = [\lambda_{jk}]_{p\times q}$. λ_{jk} is coefficient linking measured variable y_j
(2) ε and δ are residual error
(3) $\hat{\lambda}_{jk}$ is high, indicating strong association between y_j and ω_k
(4) $\hat{\Phi}_{ij}$ is high, indicating ξ_i and ξ_j are high correlated
(5) $\eta = 0.72\xi_1 + 0.1\xi_2 + 0.04\xi_3$, indicating the effect of ξ_1 is the most important

* Linear SEM: $y = \mu + \Lambda\omega + \varepsilon$, $\eta = \Gamma\xi + \delta$ (Basic assumption)
$\Rightarrow \eta_i \sim N(0, \Gamma\Phi\Gamma^T + \psi_\delta)$. Ident: Λ is non-overlapping and fixed on some value.
* SEM with fixed covariates: $y = Ac + \Lambda\omega + \varepsilon$, $\eta = Bd + \Pi\eta + \Gamma\xi + \delta$
$\Rightarrow \eta_i \sim N((I-\Pi)^{-1}Bd, (I-\Pi)^{-1}\Gamma\Phi\Gamma^T(I-\Pi)^{-T} + \psi_\delta)$, $\omega_i \sim N(\cdot,\cdot)$, $y_i \sim N(\cdot|\cdot)$
Ident. Same with the above and $\Pi_{ii} = 0$
* non-linear SEM: $y = \mu + \Lambda\omega + \varepsilon$, $\eta = \Pi\eta + \Gamma F(\xi) + \delta$
$\Rightarrow F(\xi)_{t\times1}$ nonzero, known, linearly indepentant, differentiable
y don't follow normal, $Ey_k = \mu_k + \Lambda_{k\eta}^T[(I-\Pi)^{-1}\Gamma]EF(\xi)$ Λ_k^T kth row of Λ^T

Bayes Method Advantages & Disadvan.

Adv. traditional approach: ① covariance structural analysis — under slightly more complex situation. It's not efficient and encounter theoretical and computational problems. In nonlinear SEM, outcome latent variables and y_i are not normally distributed. It's not suitable for nonliear methods ② some unnatural methods, eg. product indicator method produce inferior result.

Bayes method:
① utilizing useful prior information for achieving better results.
② for situations without accurate prior info, use noninformative prior
③ the accuracy of Bayesian estimates is close to that of MLE
④ Baysian method has potential to produce reliable results even with samples ⑤ Bayesian approach, such as Bayes factor. give more flexibility and natural tools for model comparison.

Disadv: Based on prior; need a good prior, choosen through experience

conjugate prior. $[y_i|\mu] \sim N(\mu, \sigma^2)$
① σ^2 known: $\mu \sim N(\mu_0, \tau_0^2) \Rightarrow [\mu|Y] \sim N(\hat{\mu}, \hat{\sigma}^2)$ normal-inverse-gamma
② μ known: $\sigma^2 \sim IG(\alpha, \beta) \Rightarrow [\sigma^2|Y] \sim IG(\cdot,\cdot)$
③ both unknown: $p(\mu, \sigma^2) = p(\sigma^2)p(\mu|\sigma^2) \sim N\text{-}IG \Rightarrow p(\mu, \sigma^2|Y) \sim N\text{-}IG$
④ non-informative prior, use $\frac{n}{3}$ + noninfo prior $\xrightarrow{BM}$ good prior,
⑤ in application, then use $\frac{n}{3}$ + good prior $\xrightarrow{BM}$ $\hat{\theta}$

1 choose prior
1. high conf: $\Lambda_{0k} \approx \Lambda_k$, small $H_{0yk} = 0.5I$
otherwise $\Lambda_{0k} = 0$, large $H_{0yk} = 10I$
2. $\Lambda_k^T\omega_i$ is good predictor of y_{ik}, $\psi_{\varepsilon k}$ has small E and Var ↓
3. High conf: $\Phi_0 \approx \Phi \Rightarrow R_0^{-1} = (\rho_0 - q_2 - 1)\Phi_0$
4. Λ_{0k}, $\Lambda_{0\omega k}$, Φ important that link estimates of Λ_k $\Lambda_{\omega k}$, Φ tgt
5. Source: ① expert ② Small data: noninformative
$p(\Lambda, \psi_\varepsilon) \propto \prod_{k=1}^p \psi_{\varepsilon k}^{-1}$, $p(\Lambda_\omega, \psi_\delta) \propto \prod_{k=1}^{q_1} \psi_{\delta k}^{-1}$, $p(\Phi) \propto |\Phi|^{-\frac{(q_2+1)}{2}}$
③ large data: D_0 + noninfo $\to p(\theta|D_0)$ as prior
$D\backslash D_0 + p(\theta|D_0) \to p(\theta|D)$
④ moderat data: D + noninfo $\to p(\theta|D)$ as prior
$D + p(\theta|D) \to p(\theta|D)$

$\theta \sim Gamma(\alpha, \beta)$ $p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$ $\mu = \alpha/\beta$ $\sigma^2 = \alpha/\beta^2$
$\theta \sim IG(\alpha, \beta)$ $p(\theta) = \cdots \theta^{-(\alpha+1)}e^{-\beta/\theta}$ $\mu = \beta/(\alpha-1)$ $\sigma^2 = \beta^2/(\alpha-1)^2(\alpha-2)$
$W \sim W_q(R_0, \rho_0)$ $p(W) \propto |W|^{(\rho_0-q-1)/2} \times \exp\{-\frac{1}{2}tr(R_0^{-1}W)\}$
$W \sim IW_q(R_0^{-1}, \rho_0)$ $p(W) \propto |W|^{-(\rho_0-q-1)/2} \times \exp\{-\frac{1}{2}tr(R_0^{-1}W^{-1})\}$
$EW = \rho_0 R_0$ $EW^{\cdot} = R_0^{-1}/(\rho_0 - q - 1)$.

For model $y_i = \mu + \Lambda\omega_i + \epsilon_i$,
$\eta_i = Bd_i + \Pi\eta_i + \Gamma F(\xi_i) + \delta_i$,

Assumption: ① $\varepsilon_i \overset{iid}{\sim} N(0, \Psi_\varepsilon)$ diag ② $\delta_i \overset{iid}{\sim} N(0, \Psi_\delta)$ diag ③ $\xi_i \overset{iid}{\sim} N(0, \Phi)$ general ④ $\delta_i \perp\!\!\!\perp \xi_i$, $\varepsilon_i \perp\!\!\!\perp \delta_i$, $\varepsilon_i \perp\!\!\!\perp \omega_i$
($\Lambda_k \sim N(\Lambda_{0k}, H_{0k})$)
Prior: (1) $\mu \sim N(\mu_0, \Sigma_0)$, $\psi_{\varepsilon k} \sim IG(\alpha_{0\varepsilon k}, \beta_{0\varepsilon k})$, $\Lambda_k|\psi_{\varepsilon k} \sim N(\Lambda_{0k}, \psi_{\varepsilon k}H_{0yk})$,
(2) Rewrite $\eta_i = \Lambda_\omega G(\omega_i) + \delta_i$, where $\Lambda_\omega = (B, \Pi, \Gamma)$, $G(\omega_i) = (d_i^T, \eta_i^T, F(\xi_i)^T)^T$, then $\Phi \sim IW_{q_2}(R_0^{-1}, \rho_0)$. $\psi_{\delta k} \sim IG(\alpha_{0\delta k}, \beta_{0\delta k})$,
$\Lambda_{\omega k}|\psi_{\delta k} \sim N(\Lambda_{0\omega k}, \psi_{\delta k}H_{0\omega k})$,

3. sample for posterior (Metropolis-Hastings algorithm)
(1) Data augmentation: $p(\theta|Y) \to p(\theta, \Omega|Y)$
(2) Gibbs sampling: ① initiate $\theta^{(0)}$ & $\Omega^{(0)}$
② $\theta^{(j+1)} \leftarrow p(\theta|\Omega^{(j)}, Y)$
$\Omega^{(j+1)} \leftarrow p(\Omega|\theta^{(j+1)}, Y)$
(3) Check Convergence ① Plot convergence curves from distinct initials ② Estimated potential scale reduction (EPSR) value < 1.2 $\hat{R}^{\frac{1}{2}} = [\widehat{Var}\theta/W]^{\frac{1}{2}}$
$\bar{\theta}_{\cdot k} = \frac{1}{n}\sum\theta_{jk}$ $\bar{\theta}_{\cdot\cdot} = \frac{1}{K}\sum\bar{\theta}_{\cdot k}$
$S_k^2 = \frac{1}{n-1}\sum_{j=1}^n(\theta_{jk} - \bar{\theta}_{\cdot k})^2$
$W = \frac{1}{K}\sum_{k=1}^K S_k^2$ $B = \frac{n}{K-1}\sum_{k=1}^K(\bar{\theta}_{\cdot k} - \bar{\theta}_{\cdot\cdot})^2$ $\widehat{Var}\theta = \frac{n-1}{n}W + \frac{1}{n}B$
(4) Statistical Inference
$\hat{\theta} = \frac{1}{T}\sum_{t=1}^T\theta^{(t)}$ $\widehat{Var}(\theta|Y) = \frac{1}{T-1}\sum_{t=1}^T(\theta^{(t)} - \hat{\theta})(\theta^{(t)} - \hat{\theta})^T$
$\hat{\omega}_i = \frac{1}{T}\sum\omega_i^{(t)}$
(5) Sensitivity analysis: priors for $\Lambda, \Lambda_\omega, \psi_\varepsilon, \psi_\delta, \Phi$
① different priors ② ad hoc priors

Model Comparison
1. Bayes Factor $B_{10} = \frac{P(Y|M_1)}{P(Y|M_0)} = \frac{P(Y|M_1)/P(Y|M_a)}{P(Y|M_0)/P(Y|M_a)} = B_{1a}/B_{0a}$
① $p(\Omega, \theta|Y, t) = \frac{1}{z(t)}p(Y, \Omega, \theta|t)$, $z(t) = p(Y|t)$
② $\frac{d\log z(t)}{dt} = \int\frac{1}{z(t)}\cdot\frac{d}{dt}p(Y, \Omega, \theta|t)d\Omega d\theta = E_{\Omega,\theta}[\frac{d}{dt}\log p(Y, \Omega, \theta|t)]$
$U(Y, \Omega, \theta|t) = \frac{d}{dt}\log p(Y, \Omega|\theta, t)$
$\Rightarrow \log B_{10} = \log\frac{z(1)}{z(0)} = \int_0^1\frac{d}{dt}\log z(t)dt = \int_0^1 E_{\Omega,\theta}[U(\cdots|t)]dt$
② $\widehat{\log B_{10}} = \frac{1}{2}\sum_{s=0}^S(t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)})$
$\bar{U}_{(s)} = \frac{1}{J}\sum_{j=1}^J U(Y, \Omega^{(j)}, \theta^{(j)}, t_{(s)})$ $S=20$, $J=1000$
Ex M_t: $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + (1-t)\gamma_{12}\xi_{i1}\xi_{i2} + t\gamma_{22}\xi_{i2}^2 + \delta_i$
For each $t = t_{(s)}$, $\{(\Omega^{(j)}, \theta^{(j)}), j = 1, \cdots, J\}$ ↙ N.
$U(Y, \Omega^{(j)}, \theta^{(j)}, t_{(s)}) = \frac{d}{dt}(\log p(Y, \Omega, \theta, t))|_{t=t_{(s)}}$
$= \sum_{i=1}^n\{\eta_i - \cdots - t\gamma_{22}\xi_{i2}^2\}(-\gamma_{12}\xi_{i1}\xi_{i2} + \gamma_{22}\xi_{i2}^2)/\psi_\delta$ (residual; $\frac{d}{dt}\cdots$; Σ)
Calculate $\bar{U}_{(s)}$, $s = 0, \cdots, S+1$, $\Rightarrow \widehat{\log B_{10}}$
2. DIC (Deviance Information Criterion) smaller better
$DIC = -4E_{\theta_k}[\log p(Y|\theta_k, M_k)|Y] + 2\log p(Y|\tilde{\theta}_k)$ ← Bayesian E.
$\approx \frac{1}{J}\sum_{j=1}^J\log(Y|\theta_k^{(j)}, M_k)$
* mixture model does not give DIC values in WinBUGs
* $\Delta DIC < 5$, then we should not report who is better.

---

3. L_v-measure 1. means of posterior predictive distribution 2. variances
$L_v(Y) = \sum_{i=1}^n tr[Cov(y_i^{rep}|Y)] + v\sum_{i=1}^n tr\{[E(y_i^{rep}|Y) - y_i][E(y_i^{rep}|Y) - y_i]^T\}$
$Cov(y_i^{rep}|Y) = E[Cov(y_i^{rep}|\Omega, \theta)|Y] + Cov[E(y_i^{rep}|\Omega, \theta)|Y]$
$E[E(y_i^{rep}|\Omega, \theta)|Y] = \frac{1}{J}\sum_{j=1}^J m_i^{(j)}$
$\widehat{Cov}(y_i^{rep}|Y) = \frac{1}{J}\sum_{j=1}^J[\Lambda_\eta^{(j)}(\Pi_0^{(j)})^{-1}\Psi_\delta^{(j)}(\Lambda_\eta^{(j)}(\Pi_0^{(j)})^{-1})^T + \Psi_\varepsilon^{(j)}] + \frac{1}{J}\sum_{j=1}^J m_i^{(j)}m_i^{(j)T} - (\frac{1}{J}\sum_{j=1}^J m_i^{(j)})(\frac{1}{J}\sum_{j=1}^J m_i^{(j)})^T$
$m_i^{(j)} = \mu^{(j)} + \Lambda_\eta^{(j)}(\Pi_{(0)}^{(j)})^{-1}[\Gamma^{(j)}F(\xi_i^{(j)})] + \Lambda_\xi^{(j)}\xi_i^{(j)}$
4. PP p-value (Posterior Predictive p-value) complex SEM (nonlinear) difficult to def a saturated model. previous χ
$p_B(Y) = Pr\{D(Y^{rep}|\theta, \Omega) \ge D(Y|\theta, \Omega)|Y, M_0\}$
≈ 0.5, $M_0 \sim \chi^2(d^*)$ close to 0.5 √ H_0
$\Rightarrow p_B(Y) = \int Pr(\chi^2(d^*) \ge D(Y|\theta, \Omega))p(\theta, \Omega|Y)d\theta d\Omega \approx \frac{1}{J}\sum_{j=1}^J Pr\{\chi^2(d^*) \ge D(Y|\theta^{(j)}, \Omega^{(j)})\}$.
5. Residual analysis ME: $\hat{\varepsilon}_i = y_i - \hat{A}c_i - \hat{\Lambda}\hat{\omega}_i$ outliers analysis
SE: $\hat{\delta}_i = (I - \hat{\Pi})\hat{\eta}_i - \hat{B}d_i - \hat{\Gamma}\hat{\xi}_i$
= or QQ plot of $\hat{\varepsilon}_{ij}$ $\hat{\delta}_{ik}$ → check the assumption of normality

*missing data $v_i = (x_i^T, y_i^T)^T$ latent variables.
MAR $\{\Omega, X_{mis}, Y_{mis}, Y_{obs}\} = \{\Omega, V_{mis}, Y_{obs}\}$
$p(V_{mis}|\theta, \alpha, \Omega, Y_{obs}, X_{obs}, Z_{obs}) = \prod_{i=1}^n p(V_{i,mis}|\theta, \omega_i)$
$[v_{i,mis}|\theta, \omega_i] \sim N(\mu_{i,mis} + \Lambda_{i,mis}\omega_i, \psi_{i,mis})$
Gibbs: $P(\cdot|\theta, \alpha, \Omega, V_{mis}, Y_{obs}, X_{obs}, Z_{obs}\}$ 14234
MANR: P(missing) ⫫ missing data. $r_{ij} = \begin{cases}1 & v_{ij} \text{ is mis} \\ 0 & v_{ij} \text{ is obs}\end{cases}$
$p(r|V, \Omega, \varphi) = \prod_{i=1}^n\prod_{j=1}^p[P(r_{ij}=1|v_i, \omega_i, \varphi)]^{r_{ij}}[1 - P(\cdots)]^{1-r_{ij}}$
$p(\theta, \varphi|V_{obs}, r) \propto \{\prod_{i=1}^n\int_{\omega_i, v_{i,mis}}p(v_i|\omega_i, \theta) \times p(r_i|v_i, \omega_i, \varphi)p(\omega_i|\theta)d\omega_i dv_{i,mis}\}p(\theta, \varphi)$
$m(v_i, \omega_i, \varphi) = logit\{P(r_i|v_i, \omega_i, \varphi)\} = \varphi^T e_i$ $e_i = [1, v_{i1}, \cdots, v_{ip}, \omega_{i1}, \cdots, \omega_{iq}]^T$
Gibbs: $P(\cdot|\Omega, V_{mis}, \theta, \varphi, V_{obs}, r)$
$p(\Omega|\cdot) = P(\Omega|V, \theta, \varphi, r)$; $P(V_{mis}|\cdot)$
$p(\varphi|\cdot) = P(\varphi|V, \Omega, \theta, r)$; $p(\theta|\cdot) = p(\theta|V, \Omega)$
Notation, $Y = \{y_i\}$, $D = \{d_i\}$, $X = \{x_i\}$, $Z = \{z_i\}$, $\Omega = \{\omega_i\}$
$y_i = \{y_{oi}, y_{mi}\}$, $d_i = \{d_{oi}, d_{mi}\}$, $x_i = \{x_{oi}, x_{mi}\}$, $Y_m = \{y_{mi}\}$, $D_m = \{d_{mi}\}$ $X_m = \{x_{mi}\}$
$Y_o = \{y_{oi}\}$, $D_o = \{d_{oi}\}$, $X_o = \{x_{oi}\}$, $R^y = \{r_i^y\}$, $R^d = \{r_i^d\}$, $R^x = \{r_i^x\}$
$F_m = \{Y_m, D_m, X_m, \Omega, Z\}$ $F_o = \{Y_o, D_o, X_o, R^y, R^d, R^x\}$
$\pi_k = \{\pi_{ik}\}$, $\pi = \{\pi_k\}_1^{K-1}$, $\tau = \{\tau_k\}_1^{K-1}$, $\tau_d = \{\tau_{kd}\}_1^K$,
$\varphi_y = \{\varphi_{ky}\}_1^K$, $\varphi_d = \{\varphi_{kd}\}_1^K$, $\varphi_x = \{\varphi_{kx}\}_1^K$
$\Rightarrow \begin{cases}\theta_s = \{\varphi_y, \varphi_d, \varphi_x, \tau, \tau_x, \tau_d\} \\ \theta_* = \{\theta, \pi, \theta_s\}\end{cases}$
$Mean(\hat{\theta}) = \frac{1}{R}\sum\hat{\theta}_r$, $MC\text{-}SD(\hat{\theta})\sqrt{\frac{1}{R-1}\sum(\hat{\theta}_r - \theta)^2}$
$EST\text{-}SD(\hat{\theta}) = \sqrt{\widehat{Var}(\theta|V)}$ $RMS(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^R(\hat{\theta}_r - \theta)^2}$
$SE/SD(\hat{\theta}) = \frac{EST\text{-}SD(\hat{\theta})}{MC\text{-}SD(\hat{\theta})}$
Example $\vec{v}_i = (z_{i1}, z_{i2}, z_{i3}, x_{i4}, \dots, x_{i9})^T$ (dich; binary; Bernoulli (P_{ik}))
$x_{ik} \propto \exp\{x_{ik}\vartheta_{ik} - \log(1 + e^{\vartheta_{ik}})\}$ ($b(\vartheta_{ik})$) $\vartheta_{ik} = \log(\frac{P_{ik}}{1 - P_{ik}})$
$y_{ik} = \mu_k + \Lambda_k^T\omega_i + \varepsilon_i$ $\vartheta_{ik} = \mu_k + \Lambda_k^T\omega_i$ $\psi_{\varepsilon k} = 1$ fixed

---

* Ordered categorical variables. $\Omega \to Y \to Z$
Assumption $y \sim N \to z \sim OrderCat$
by $z = m$ if $\alpha_m \le y < \alpha_{m+1}$ $m = 0, \cdots, b$
$-\infty = \alpha_{k,0} < \alpha_{k,1} < \cdots < \alpha_{k,b_k} < \alpha_{k,b_k+1} = +\infty$
Identification conditions for $k = 1, \dots, s$ ...
① fix some elements and nonoverlap in Λ, Π, and Γ
② indeterminacy induced by the orderCat variable. by imposing $\alpha_{k,1} = \Phi^{*-1}(f_1^*)$, $\alpha_{k,b_k} = \Phi^{*-1}(f_{k,b_k}^*)$ Φ^* cdf of z
implicitly unifing $y \sim N(0,1)$ f_1^* frequency of first class, $f_{b_k}^*$ cummulative of $z_k < b_k$
* Dichotomous variables: only one threshold.
Assumption $y_{ik} \sim N \to z_{ik} = \begin{cases}1 & \text{if } y_{ik} > 0 \\ 0 & \text{otherwise}\end{cases}$ (not ordere
$P(z_{ik} = 1|\omega_i, \mu_k, \Lambda_k, \psi_{\varepsilon k}) = \Phi([\mu_k + \Lambda_k^T\omega_i]/\psi_{\varepsilon k}^{1/2})$ no threshold needs to be estimated in Gibbs sampler.
Identification conditions ① fix and nonoverlap
② fix $\psi_{\varepsilon k} = 1$ for the covariance (Next Page)
Prior: $\alpha \perp\!\!\!\perp \theta$, $\alpha_k \perp\!\!\!\perp \alpha_l$
$p(\alpha_k) = p(\alpha_{k,2}, \dots, \alpha_{k,b_k-1}) \propto c$ noninfo.
$p(\alpha_k, Y_k|Z_k, \theta, \Omega) = p(\alpha_k|Z_k, \theta, \Omega)p(Y_k|\alpha_k, Z_k, \theta, \Omega)$
$p(\alpha_k|Z_k, \theta, \Omega) \propto \prod_{i=1}^n[\Phi^*\{\psi_{yk}^{-\frac{1}{2}}(\alpha_{k,z_{ik}+1} - \mu_{yk} - \Lambda_{yk}^T\omega_i)\} - \Phi^*\{\psi_{yk}^{-\frac{1}{2}}(\alpha_{k,z_{ik}} - \mu_{yk} - \Lambda_{yk}^T\omega_i)\}]$
$[y_{ik}|\alpha_k, Z_k, \theta, \Omega] \sim N(\mu_{yk} + \Lambda_{yk}^T\omega_i, \psi_{yk})I_{(\alpha_{k,z_{ik}}, \alpha_{k,z_{ik}+1})}(y_{ik})$ EFD
$p(\alpha_k, Y_k|Z_k, \theta, \Omega) \propto \prod_{i=1}^n[\Phi^*\{\psi_{yk}^{-\frac{1}{2}}(y_{ik} - \mu_{yk} - \Lambda_{yk}^T\omega_i)\}I_{(\alpha_{k,z_{ik}}, \alpha_{k,z_{ik}+1})}(y_{ik})$.
Gibbs sampler
(1) $\Omega^{(j+1)} \leftarrow P(\Omega|\theta^{(j)}, \alpha^{(j)}, Y^{(j)}, X, Z)$
(2) $\theta^{(j+1)} \leftarrow P(\theta|\Omega^{(j+1)}, \alpha^{(j)}, Y^{(j)}, X, Z)$
(3) OrdCat $\alpha^{(j+1)}, Y^{(j+1)} \leftarrow P(\alpha, Y|\theta^{(j+1)}, \Omega^{(j+1)}, X, Z)$
Dichot $[y_{ik}|\theta, \omega_i, z_i] \sim \begin{cases}N(\mu_k + \Lambda_k^T\omega_i, 1)1_{(-\infty,0]}(y_{ik}), & z_{ik} = 0 \\ N(\mu_k + \Lambda_k^T\omega_i, 1)1_{(0,+\infty)}(y_{ik}), & z_{ik} = 1\end{cases}$
* Exponential family Assumption count; binary (ordered)
$p(y_{ik}|\omega_i) = \exp\{[y_{ik}\vartheta_{ik} - b(\vartheta_{ik})]/\psi_{\varepsilon k} + c_k(y_{ik}, \psi_{\varepsilon k})\}$
$E(y_{ik}|\omega_i) = \dot{b}(\vartheta_{ik})$ $Var(y_{ik}|\omega_i) = \psi_{\varepsilon k}\ddot{b}(\vartheta_{ik})$
$\vartheta_{ik} = A_k^T c_{ik} + \Lambda_k^T\omega_i$ $\eta_i = Bd_i + \Pi\eta_i + \Gamma F(\xi_i) + \delta_i$
Identification: same with OrdCat.
BF in EFD: $D_o = (X, Z)$. $D_m = (Y, \Omega)$, $\theta^* = (\theta, \alpha)$
$U(\theta^*, D_m, D_o, t) = \frac{d}{dt}(\log p(D_m, D_o|\theta^*, t)$
$\bar{U}$ & $\widehat{\log B_{10}}$ are same with the standard one.
Posterior $p(\Omega|Y, \theta) \propto \prod_{i=1}^n p(y_i|\omega_i, \theta)p(\eta_i|\xi_i, \theta)p(\xi_i|\theta)$
$p(\omega_i|y_i, \theta) \propto \exp\{\sum_{k=1}^p[y_{ik}\vartheta_{ik} - b(\vartheta_{ik})]/\psi_{\varepsilon k} - \frac{1}{2}[\eta_i - Bd_i - \Pi\eta_i - \Gamma F(\xi_i)]^T\Psi_\delta^{-1}[\cdots] + \xi_i^T\Phi^{-1}\xi_i]$
$p(A_k|Y, \Omega, \Lambda_k, \psi_{\varepsilon k}) \propto \exp\{\sum_{i=1}^n\frac{y_{ik}\vartheta_{ik} - b(\vartheta_{ik})}{\psi_{\varepsilon k}} - \frac{1}{2}(A_k - A_{0k})^T H_k^{-1}\cdots$
$p(\psi_{\varepsilon k}|Y, \Omega, A_k, \Lambda_k) \propto \psi_{\varepsilon k}^{-(\frac{n}{2} + \alpha_{0\varepsilon k} - 1)}\exp\{\sum_i \cdots + c_k(y_{ik}, \psi_{\varepsilon k}) - \frac{\beta_{0\varepsilon k}}{\psi_{\varepsilon k}}\}$
$p(\Lambda_k|Y, \Omega, A_k, \psi_{\varepsilon k}) \propto \exp\{\sum_{i=1}^n\frac{y_{ik}\vartheta_{ik} - b(\vartheta_{ik})}{\psi_{\varepsilon k}} - \frac{1}{2}\psi_{\varepsilon k}^{-1}(\Lambda_k - \Lambda_{0k})^T H_{0yk}^{-1}\cdots)$
SEM Bayesian is same.
① large sample size needed ② $\sqrt{\widehat{Var}(\theta|V)}$ overest true SD ③ incorrectly treat dichotomous as binary or vice versa would produce misleading result. ($\frac{SE}{SD} > 1$).

$BIC_k = -2\log p(Y|\tilde{\theta}_1, M_1) + d_k\log n$
$BIC_{10} = -2\log B_{10} = 2\log B_{01}$
$AIC_k = -2\log p(Y|\tilde{\theta}_k, M_k) + 2d_k$
① BIC tends to favor simpler models.
② M_k is selected if AIC_k, BIC_k small
③ $2\log B_{10} < 0 \Rightarrow$ support M_0.

* Hierarchical Data — come from a number of diff groups with known hierarchical structure!
L1 (within group) $u_{gi} = v_g + \Lambda_{1g}\omega_{1gi} + \varepsilon_{1gi}$ $g = 1, \cdots, G$, $i = 1, \cdots, N_g$
L2 (between group) $v_g = \mu + \Lambda_2\omega_{2g} + \varepsilon_{2g}$ $u_{gi} = \begin{bmatrix}x_{gi} \\ y_{gi}\end{bmatrix} \to z_{gi}$
$\Rightarrow u_{gi} = \mu + \Lambda_2\omega_{2g} + \varepsilon_{2g} + \Lambda_{1g}\omega_{1gi} + \varepsilon_{1gi}$
$\eta_{1gi} = \Pi_{1g}\eta_{1gi} + \Gamma_{1g}F_1(\xi_{1gi}) + \delta_{1gi}$
$\eta_{2g} = \Pi_2\eta_{2g} + \Gamma_2F_2(\xi_{2g}) + \delta_{2g}$
Assumption within group params are invariant/shared. coeff & errors are distinct.
Example M_1: $v_g = \mu + \Lambda_2^1\omega_{2g} + \varepsilon_{2g}$ $\omega_{2g} \sim N(0, \Phi_2)$ FA
M_2: $v_g = \mu + \Lambda_2^2\omega_{2g} + \varepsilon_{2g}$
$\eta_{2g} = \Pi_2^2\eta_{2g} + \Gamma_2^2F_2(\xi_{2g}) + \delta_{2g}$
M_a: $v_g = \mu + \varepsilon_{2g}$
link model M_{t1a}: $u_{gi} = \mu + t\Lambda_2^1\omega_{2g} + \varepsilon_{2g} + \Lambda_1\omega_{1gi} + \varepsilon_{1gi}$
$\omega_{2g} \sim N(0, \Phi_2)$ $\eta_{1gi} = \Pi_1\eta_{1gi} + \Gamma_1F_1(\xi_{1gi}) + \delta_{1gi}$
M_{t2a}: $u_{gi} = \mu + t\Lambda_2^2\omega_{2g} + \Lambda_1\omega_{1gi} + \varepsilon_{1gi}$
$\Rightarrow \widehat{\log B_{12}} = \widehat{\log B_{1a}} - \widehat{\log B_{2a}}$

* Multisample data small G large N_g Obs_{gi} ⫫ $i = 1, \dots$
target investigate the similarity or diff among groups
test: H_0: params invariant among groups.
$v_i^{(g)} = \mu^{(g)} + \Lambda^{(g)}\omega_i^{(g)} + \varepsilon_i^{(g)}$ $\sim N(0, \Psi_\varepsilon^{(g)})$
$\begin{bmatrix}x_i^{(g)} \\ y_i^{(g)}\end{bmatrix} \to z_i^{(g)}$ $\sim N(0, \Phi^{(g)})$
$\eta_i^{(g)} = \Pi^{(g)}\eta_i^{(g)} + \Gamma^{(g)}F(\xi_i^{(g)}) + \delta_i^{(g)} \sim N(0, \Psi_\delta^{(g)})$
Assumption parameters have the same dimension among group
identification select the first group as reference
① fix $\alpha_{k,1}^{(g)}$ & $\alpha_{k,b_k}^{(g)}$ among groups sharing scale
eg. $\alpha_{k,b}^{(g)} = \alpha_{k,b}^{(1)}$, $k = 1, \dots, s$. $b = 1, \dots, b_k$

Prior: Let $\Lambda_{\omega k}^T$ and $\Lambda_{\omega k}^{(g)T}$ be the kth rows of Λ_ω and $\Lambda_\omega^{(g)}$, respectively. The prior distributions of Λ_ω are given by:

(I) If $\Lambda_\omega^{(1)} = \cdots = \Lambda_\omega^{(G)} = \Lambda_\omega$, $\Lambda_{\omega k} \overset{D}{=} N[\Lambda_{0\omega k}, H_{0\omega k}]$, constraint
(II) If $\Lambda_\omega^{(1)} \ne \cdots \ne \Lambda_\omega^{(G)}$, $\Lambda_{\omega k}^{(g)} \overset{D}{=} N[\Lambda_{0\omega k}^{(g)}, H_{0\omega k}^{(g)}]$. non constraint.

Similarly, let $\psi_{\delta k}^{(g)}$ be the kth diagonal element of $\Psi_\delta^{(g)}$. The prior distributions of $\psi_{\delta k}^{(g)-1}$ $\mu^{(g)}$, and $\Phi^{(g)}$ are

$$Gamma[\alpha_{0\delta k}^{(g)}, \beta_{0\delta k}^{(g)}], \quad \mu^{(g)} \overset{D}{=} N[\mu_0^{(g)}, \Sigma_0^{(g)}], \quad \Phi^{(g)-1} \overset{D}{=} W_{q_2}[R_0^{(g)}, \rho_0^{(g)}].$$

model comparison

Let M_k and M_h be two nested or non-nested competing models (hypotheses), M_k and M_h can then be compared using the Bayes factor:

$$B_{kh} = \frac{p(X, Z|M_k)}{p(X, Z|M_h)}.$$

In computing $\log B_{kh}$ through path sampling, searching for a good path to link M_k and M_h is a crucial step. As an illustrative example, suppose that the competing models M_1 and M_2 are defined as follows. For $g = 1, 2$, $i = 1, \cdots, N_g$,

$$M_1: \quad v_i^{(g)} = \mu^{(g)} + \Lambda\omega_i^{(g)} + \epsilon_i^{(g)}, \quad \eta_i^{(g)} = \Gamma^{(g)}F(\xi_i^{(g)}) + \delta_i^{(g)};$$

H_0: $\Lambda_1 = \cdots = \Lambda_G$

$$M_2: \quad v_i^{(g)} = \mu^{(g)} + \Lambda^{(g)}\omega_i^{(g)} + \epsilon_i^{(g)}, \quad \eta_i^{(g)} = \Gamma F(\xi_i^{(g)}) + \delta_i^{(g)}.$$

H_1: $\Gamma_1 = \cdots = \Gamma_G$.

$$M_a: \quad v_i^{(g)} = \mu^{(g)} + \epsilon_i^{(g)}, \quad g = 1, 2, \; i = 1, \cdots, N_g.$$

The link model M_{ta} for linking M_1 and M_a is define below:

$$M_{ta1}: \quad v_i^{(g)} = \mu^{(g)} + t\Lambda\omega_i^{(g)} + \epsilon_i^{(g)}, \quad \eta_i^{(g)} = \Gamma^{(g)}F(\xi_i^{(g)}) + \delta_i^{(g)}, \quad g = 1, 2, \; i = 1, \cdots, N_g.$$

When $t = 1$, M_{ta1} reduces to M_1, and when $t = 0$, M_{ta1} reduces to M_a. The parameter vector θ in M_{ta1} contains $\mu^{(1)}$, $\mu^{(2)}$, Λ, $\Psi_\epsilon^{(1)}$, $\Psi_\epsilon^{(2)}$, $\Gamma^{(1)}$, $\Gamma^{(2)}$, $\Phi^{(1)}$, $\Phi^{(2)}$, $\Psi_\delta^{(1)}$, and $\Psi_\delta^{(2)}$.
The link model M_{ta2} for linking M_2 and M_a is defined as follows:

$$M_{ta2}: \quad v_i^{(g)} = \mu^{(g)} + t\Lambda^{(g)}\omega_i^{(g)} + \epsilon_i^{(g)}, \quad \eta_i^{(g)} = \Gamma F(\xi_i^{(g)}) + \delta_i^{(g)}.$$

Clearly, when $t = 1$ and 0, M_{ta2} reduces to M_2 and M_a, respectively. The parameter vector in M_{ta2} contains $\mu^{(1)}$, $\mu^{(2)}$, $\Lambda^{(1)}$, $\Lambda^{(2)}$, $\Psi_\epsilon^{(1)}$, $\Psi_\epsilon^{(2)}$, Γ, $\Phi^{(1)}$, $\Phi^{(2)}$, $\Psi_\delta^{(1)}$, and $\Psi_\delta^{(2)}$.

We first compute $\log B_{1a}$ and $\log B_{2a}$, and then obtain $\log B_{12}$ via the following equation:

$$\log B_{12} = \log\frac{p(X, Z|M_1)/p(X, Z|M_a)}{p(X, Z|M_2)/p(X, Z|M_a)} = \log B_{1a} - \log B_{2a}.$$

where $\bar{Y}_k = \sum_{i:w_i=k}(y_i - \Lambda_k\omega_i)/n_k$, with $\sum_{i:w_i=k}$ denotes the summation with respect to those i such that $w_i = k$, and

$$a_{ykm} = A_{ykm}(H_{0ykm}^{-1}\Lambda_{0km} + \Omega_k\tilde{Y}_{km}), \quad A_{ykm} = (H_{0ykm}^{-1} + \Omega_k\Omega_k^T)^{-1},$$
$$\beta_{\epsilon km} = \beta_{0\epsilon k} + [\tilde{Y}_{km}^T\tilde{Y}_{km} - a_{ykm}^T A_{ykm}^{-1}a_{ykm} + \Lambda_{0km}^T H_{0ykm}^{-1}\Lambda_{0km}]/2,$$
$$a_{\delta kl} = A_{\omega kl}(H_{0\omega kl}^{-1}\Lambda_{0\omega kl} + \Omega_k\Omega_{1kl}), \quad A_{\omega kl} = (H_{0\omega kl}^{-1} + \Omega_k\Omega_k^T)^{-1},$$
$$\beta_{\delta kl} = \beta_{0\delta k} + [\Omega_{1kl}^T\Omega_{1kl} - a_{\delta kl}^T A_{\omega kl}^{-1}a_{\delta kl} + \Lambda_{0\omega kl}^T H_{0\omega kl}^{-1}\Lambda_{0\omega kl}]/2,$$

in which $\tilde{Y}_{km}^T$ is the mth row of $\tilde{Y}_k$, and $\tilde{Y}_k$ is a matrix whose columns are equal to the columns of Y_k minus μ_k, and Ω_{1kl}^T is the lth row of Ω_{1k}. The computational burden in

---

Differences between mixture, multisapmle, and hierarchical data:

1. K is usually small in mixture and multisample data, but is large for hierarchical data. The number of observations within each group is large for mixture and multisample data, but is relatively small for hierarchical data.
2. The group membership of each observation is unknown for mixture data, but is specified for hierarchical and multisample data.
3. The observations in each group are independent for mixture and multisample data, but dependent for hierarchical data because individuals within a group share certain common influential factors.

Mixture SEM

· Let y_i be a $p\times1$ random vector corresponding to the i th observation, and its distribution is $f(y_i|\theta) = \sum_{k=1}^K \pi_k f_k(y_i|\mu_k, \theta_k)$, $i = 1, \cdots, n$ where K is a given integer, π_k is the unknown mixing proportion such that $\pi_k > 0$ and $\pi_1 + \cdots + \pi_K = 1$, $f_k(y_i|\mu_k, \theta_k)$ is the multivariate normal density function. For the k-th component, the measurement equation of the model is given by $y_i = \mu_k + \Lambda_k\omega_i + \epsilon_i$, $\eta_i = \Pi_k\eta_i + \Gamma_k\xi_i + \delta_i$.
· **Identifiability :** (1) If $\mu_{1,1} < \cdots < \mu_{K,1}$ are well separated, we impose the ordering $\mu_{1,1} < \cdots < \mu_{K,1}$ for solving the label switching problem. (2) If $\mu_{1,1} < \cdots < \mu_{K,1}$ are close to each other, we use random permutation sampler. (3) For each $k = 1, \cdots K$, the SEM is identified by fixing appropriate elements in Λ_k, Π_k, and/or Γ_k at preassigned values.
· (Random Permutation) Let $\psi = (\Omega, W, \theta)$, the permutation sampler for generating ψ from the posterior $p(\psi|Y)$ is implemented as follows : (1) Generate $\tilde{\psi}$ from the unconstrained posterior $p(\psi|Y)$ using standard Gibbs sampling steps; (2) Select some permutation $\rho(1), \cdots, \rho(K)$ and define $\psi = \rho(\tilde{\psi})$ from $\tilde{\psi}$ by reordering the labeling through this permutation : $(\theta_1, \cdots, \theta_K) := (\theta_{\rho(1)}, \cdots, \theta_{\rho(K)})$, and $W = (w_1, \cdots, w_n) := (\rho(w_1), \cdots, \rho(w_n))$ solve label switching

model comparison:
Consider the following competing models:

$$M_1: \quad Y|\theta, \pi \overset{D}{=} \sum_{k=1}^K \pi_k f_k(Y|\mu_k, \Sigma_k), \quad (15)$$

$$M_0: \quad Y|\theta, \pi^* \overset{D}{=} \sum_{k=1}^c \pi_k^* f_k(Y|\mu_k, \Sigma_k), \quad (16)$$

where $1 \le c < K$. They can be linked up by a path M_t as follows:

$$M_t: \quad [y|\theta, \pi, t] \overset{D}{=} [\pi_1 + (1-t)a_1(\pi_{c+1} + \cdots + \pi_K)]f_1(y|\mu_1, \Sigma_1) + \cdots + [\pi_c + (1-t)a_c(\pi_{c+1} + \cdots + \pi_K)]f_c(y|\mu_c, \Sigma_c) + t\pi_{c+1}f_{c+1}(y|\mu_{c+1}, \Sigma_{c+1}) + \cdots + t\pi_K f_K(y|\mu_K, \Sigma_K),$$

$$\log p(Y, \Omega|\theta, t) = \sum_{i=1}^n \log\left\{\sum_{k=1}^c[\pi_k + (1-t)a_k\sum_{h=c+1}^K\pi_h] \times f_k(y_i, \omega_i|\mu_k, \Sigma_k) + \sum_{k=c+1}^K t\pi_k f_k(y_i, \omega_i|\mu_k, \Sigma_k)\right\}.$$

By differentiation with respect to t, we have

$$U(Y, \Omega, \theta, t) = \sum_{i=1}^n \frac{-\sum_{h=c+1}^K\pi_h\sum_{k=1}^c a_k f_k(y_i, \omega_i|\mu_k, \Sigma_k) + \sum_{k=c+1}^K\pi_k f_k(y_i, \omega_i|\mu_k, \Sigma_k)}{\sum_{k=1}^c\left[\pi_k + (1-t)a_k\sum_{h=c+1}^K\pi_h\right]f_k(y_i, \omega_i|\mu_k, \Sigma_k) + \sum_{k=c+1}^K t\pi_k f_k(y_i, \omega_i|\mu_k, \Sigma_k)},$$

where $f_k(\cdot)$ is written as follows.

$$f_k(y_i, \omega_i|\mu_k, \Sigma_k) = (2\pi)^{-p/2}|\Psi_k|^{-1/2} \times \exp\left[-\frac{1}{2}(y_i - \mu_k - \Lambda_k\omega_i)^T\Psi_k^{-1}(y_i - \mu_k - \Lambda_k\omega_i)\right] \times (2\pi)^{-q_1/2}|I_{q_1} - \Pi_k||\Psi_{\delta k}|^{-1/2} \times \exp\left[-\frac{1}{2}(\eta_i - \Lambda_{\omega k}\omega_i)^T\Psi_{\delta k}^{-1}(\eta_i - \Lambda_{\omega k}\omega_i)\right] \times (2\pi)^{-q_2/2}|\Phi_k|^{-1/2}\exp\left(-\frac{1}{2}\xi_i^T\Phi_k^{-1}\xi_i\right),$$

and $\Lambda_{\omega k} = (\Pi_k, \Gamma_k)$. Thus, the Bayes factor can be estimated with

$$\bar{U}_{(s)} = J^{-1}\sum_{j=1}^J U(Y, \Omega^{(j)}, \theta^{(j)}, t_{(s)}),$$

Posterior distributions of mixture model:

Step (a): Generate $(W^{(r+1)}, \Omega^{(r+1)})$ from $p(\Omega, W|Y, \theta^{(r)})$;

Step (b): Generate $\theta^{(r+1)}$ from $p(\theta|Y, \Omega^{(r+1)}, W^{(r+1)})$;

Step (c): Reorder the label through the permutation sampler to achieve the identifiability.

As $p(\Omega, W|Y, \theta) = p(W|Y, \theta)p(\Omega|Y, W, \theta)$, Step (a) can be further decomposed into the following two steps:

Step (a1): Generate $W^{(r+1)}$ from $p(W|Y, \theta^{(r)})$;

Step (a2): Generate $\Omega^{(r+1)}$ from $p(\Omega|Y, \theta^{(r)}, W^{(r+1)})$.

(a1)
$$p(W|Y, \theta) = \prod_{i=1}^n p(w_i|y_i, \theta). \quad (7.5)$$

Moreover,

$$p(w_i = k|y_i, \theta) = \frac{p(w_i = k, y_i|\theta)}{p(y_i|\theta)} = \frac{p(w_i = k|\pi)p(y_i|w_i = k, \theta)}{p(y_i|\theta)} = \frac{\pi_k f_k(y_i|\mu_k, \theta_k)}{f(y_i|\theta)}, \quad (7.6)$$

(a2)
$$\prod_{i=1}^n p(\omega_i|y_i, w_i, \theta) = p(\Omega|Y, \theta, W) \propto p(Y|\Omega, W, \mu, \theta_y)p(\Omega|W, \theta_\omega) = \prod_{i=1}^n p(y_i|\omega_i, w_i, \mu, \theta_y)p(\omega_i|w_i, \theta_\omega). \quad (7.7)$$

Let $C_k = \Sigma_{\omega k}^{-1} + \Lambda_k^T\Psi_k^{-1}\Lambda_k$, where $\Sigma_{\omega k}$ is the covariance matrix of ω_i in the kth compo-

$$[\omega_i|y_i, w_i = k, \theta] \overset{D}{=} N[C_k^{-1}\Lambda_k^T\Psi_k^{-1}(y_i - \mu_k), C_k^{-1}].$$

(b)
$$p(\theta|W, \Omega, Y) = p(\pi, \mu, \theta_y, \theta_\omega|W, \Omega, Y) = [p(\pi)p(W|\pi)][p(\mu)p(\theta_y)p(Y|W, \Omega, \mu, \theta_y)][p(\theta_\omega)p(\Omega|W, \theta_\omega)]$$

prior:
$[\Lambda_{km}|\psi_{km}] \overset{D}{=} N[\Lambda_{0km}, \psi_{km}H_{0ykm}]$, $\psi_{km}^{-1} \overset{D}{=} Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}]$,
$[\Lambda_{\omega kl}|\psi_{\delta kl}] \overset{D}{=} N[\Lambda_{0\omega kl}, \psi_{\delta kl}H_{0\omega kl}]$, $\psi_{\delta kl}^{-1} \overset{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}]$,
$\mu_k \overset{D}{=} N[\mu_{0k}, \Sigma_{0k}]$, $\Phi_k^{-1} \overset{D}{=} W_{q_2}[R_{0k}, \rho_{0k}]$,
$[\mu_k|Y_k, \Omega_k, \Lambda_k, \Psi_k] \overset{D}{=} N[(\Sigma_0^{-1} + n_k\Psi_k^{-1})^{-1}(n_k\Psi_k^{-1}\bar{Y}_k + \Sigma_0^{-1}\mu_0), (\Sigma_0^{-1} + n_k\Psi_k^{-1})^{-1}]$,
$[\Lambda_{km}|Y_k, \Omega_k, \psi_{km}^{-1}] \overset{D}{=} N[a_{ykm}, \psi_{km}A_{ykm}]$,
$[\psi_{km}^{-1}|Y_k, \Omega_k, \mu_{km}] \overset{D}{=} Gamma[n_k/2 + \alpha_{0\epsilon k}, \beta_{\epsilon km}]$,
$[\Lambda_{\omega kl}|Y_k, \Omega_k, \psi_{\delta kl}^{-1}] \overset{D}{=} N[a_{\delta kl}, \psi_{\delta kl}A_{\omega kl}]$,
$[\psi_{\delta kl}^{-1}|Y_k, \Omega_k] \overset{D}{=} Gamma[n_k/2 + \alpha_{0\delta k}, \beta_{\delta kl}]$,
$[\Phi_k|\Omega_{2k}] \overset{D}{=} IW_{q_2}[(\Omega_{2k}\Omega_{2k}^T + R_0^{-1}), n_k + \rho_0]$,

Continue for 5020. (make up some hints)

2. Derive posterior

Example 1.

Example 2.

SEM with ordered categorical

SEM with dichotomous variables. ("True" or "False")

EFD example:

EFD: binary dicho: ordered cat.

EFD model:

Prior:

EFD) posterior

MH for $[\theta, \Omega, \alpha, \Upsilon | X, Z]$

Example of EFD: Logit regression

$\Sigma | M, Y$

$M | Y, \Sigma$ (conjugate prior)

$M | Y, \Sigma$ (non-info prior) Let $p(M) \propto c$ be the non-info prior,

$P(Y, M | \Sigma) \propto cp(Y | M, \Sigma) = c(2\pi)^{-\frac{P}{2}} |\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(Y-M)^T \Sigma^{-1}(Y-M))$

then $M | Y, \Sigma \sim N(Y, \Sigma)$

Appendix 6.5: Conditional Distributions for Multisample SEMs

The conditional distributions $[\Omega | \theta, \alpha, \mathbf{Y}, \mathbf{X}, \mathbf{Z}]$, $[\alpha, \mathbf{Y} | \theta, \Omega, \mathbf{X}, \mathbf{Z}]$, and $[\theta | \alpha, \mathbf{Y}, \Omega, \mathbf{X}, \mathbf{Z}]$

$$p[\Omega | \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}] = \prod_{g=1}^{G} \prod_{i=1}^{N_g} p(\omega_i^{(g)} | \mathbf{v}_i^{(g)}, \boldsymbol{\theta}^{(g)}),$$

where

$$p(\omega_i^{(g)} | \mathbf{v}_i^{(g)}, \boldsymbol{\theta}^{(g)}) \propto \exp\left\{ -\frac{1}{2} \left[(\mathbf{v}_i^{(g)} - \boldsymbol{\mu}^{(g)} - \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)})^T \boldsymbol{\Psi}_\epsilon^{(g)-1} (\mathbf{v}_i^{(g)} - \boldsymbol{\mu}^{(g)} - \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)}) + (\boldsymbol{\eta}_i^{(g)} - \boldsymbol{\Lambda}_\omega^{(g)} \mathbf{G}(\omega_i^{(g)}))^T \boldsymbol{\Psi}_\delta^{(g)-1} (\boldsymbol{\eta}_i^{(g)} - \boldsymbol{\Lambda}_\omega^{(g)} \mathbf{G}(\omega_i^{(g)})) + \boldsymbol{\xi}_i^{(g)T} \boldsymbol{\Phi}^{(g)-1} \boldsymbol{\xi}_i^{(g)} \right] \right\}. \quad (6.A10)$$

$$p(\boldsymbol{\alpha}, \mathbf{Y} | \cdot) = \prod_{g=1}^{G} p(\boldsymbol{\alpha}^{(g)}, \mathbf{Y}^{(g)} | \cdot) = \prod_{g=1}^{G} \prod_{k=1}^{s} p(\boldsymbol{\alpha}_k^{(g)}, \mathbf{Y}_k^{(g)} | \cdot),$$

$$p(\boldsymbol{\alpha}_k^{(g)}, \mathbf{Y}_k^{(g)} | \cdot) \propto \prod_{i=1}^{N_g} \phi\{\psi_{cgk}^{(g)-1/2} (y_{ik}^{(g)} - \boldsymbol{\mu}_{yk}^{(g)} - \boldsymbol{\Lambda}_{yk}^{(g)T} \boldsymbol{\omega}_i^{(g)})\} I_{[\alpha_{k,z_{ik}}, \alpha_{k,z_{ik}+1})}(y_{ik}^{(g)}),$$

The conditional distribution of some components in $\boldsymbol{\theta}^{(g)}$, $g = 1, \cdots, G$, under the situation without any parameter constraints are given as follows. Let $\boldsymbol{\Lambda}_k^{(g)T}$ be the kth row of $\boldsymbol{\Lambda}^{(g)}$, and $\psi_{ck}^{(g)}$ be the kth diagonal element of $\boldsymbol{\Psi}_\epsilon^{(g)}$, $\mathbf{V}_k^{*(g)}$ be the kth row of $\mathbf{V}^{*(g)} = (\mathbf{v}_1^{(g)} - \boldsymbol{\mu}^{(g)}, \cdots, \mathbf{v}_{N_g}^{(g)} - \boldsymbol{\mu}^{(g)})$, and $\Omega_2^{(g)} = (\boldsymbol{\xi}_1^{(g)}, \cdots, \boldsymbol{\xi}_{N_g}^{(g)})$. It can be shown that:

$$[\boldsymbol{\mu}^{(g)} | \boldsymbol{\Lambda}^{(g)}, \boldsymbol{\Psi}_\epsilon^{(g)}, \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{X}] \overset{D}{=} N(\mathbf{a}_\mu^{(g)}, \mathbf{A}_\mu^{(g)}),$$

$$[\boldsymbol{\Lambda}_k^{(g)} | \boldsymbol{\Psi}_\epsilon^{(g)}, \boldsymbol{\mu}^{(g)}, \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{X}] \overset{D}{=} N(\mathbf{a}_k^{(g)}, \mathbf{A}_k^{(g)}),$$

$$[\psi_{ck}^{(g)-1} | \boldsymbol{\Lambda}_k^{(g)}, \boldsymbol{\mu}^{(g)}, \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{X}] \overset{D}{=} Gamma(N_g/2 + \alpha_{0ck}^{(g)}, \beta_{ck}^{(g)}),$$

$$[\boldsymbol{\Phi}^{(g)} | \boldsymbol{\Omega}_2^{(g)}] \overset{D}{=} IW_{q_2}[(\boldsymbol{\Omega}_2^{(g)} \boldsymbol{\Omega}_2^{(g)T} + \mathbf{R}_0^{(g)-1}), N_g + \rho_0^{(g)}],$$

in which

$$\mathbf{a}_\mu^{(g)} = \mathbf{A}_\mu^{(g)} [\Sigma_0^{(g)-1} \boldsymbol{\mu}_0^{(g)} + N_g \boldsymbol{\Psi}_\epsilon^{(g)-1} (\bar{\mathbf{v}}^{(g)} - \boldsymbol{\Lambda}^{(g)} \bar{\boldsymbol{\omega}}^{(g)})], \quad \mathbf{A}_\mu^{(g)} = (\Sigma_0^{(g)-1} + N_g \boldsymbol{\Psi}_\epsilon^{(g)-1})^{-1},$$

$$\mathbf{a}_k^{(g)} = \mathbf{A}_k^{(g)} [\mathbf{H}_{0gk}^{(g)-1} \boldsymbol{\Lambda}_{0k}^{(g)} + \psi_{ck}^{(g)-1} \boldsymbol{\Omega}^{(g)} \mathbf{V}_k^{*(g)}], \quad \mathbf{A}_k^{(g)} = [\psi_{ck}^{(g)-1} \boldsymbol{\Omega}^{(g)} \boldsymbol{\Omega}^{(g)T} + \mathbf{H}_{0gk}^{(g)-1}]^{-1}$$

$$\beta_{ck}^{(g)} = \beta_{0ck}^{(g)} + [\boldsymbol{\Lambda}_k^{(g)T} \boldsymbol{\Omega}^{(g)} \boldsymbol{\Omega}^{(g)T} \boldsymbol{\Lambda}_k^{(g)} - 2\boldsymbol{\Lambda}_k^{(g)T} \boldsymbol{\Omega}^{(g)} \mathbf{V}_k^{*(g)} + \mathbf{V}_k^{*(g)T} \mathbf{V}_k^{*(g)}]/2,$$

with $\bar{\mathbf{v}}^{(g)} = \sum_{i=1}^{N_g} \mathbf{v}_i^{(g)}/N_g$ and $\bar{\boldsymbol{\omega}}^{(g)} = \sum_{i=1}^{N_g} \boldsymbol{\omega}_i^{(g)}/N_g$ are the means of $\mathbf{v}_i^{(g)}$ and $\boldsymbol{\omega}_i^{(g)}$ within the gth group.

① under $\boldsymbol{\Lambda}_k^{(1)} = \boldsymbol{\Lambda}_k^{(2)} \cdots = \boldsymbol{\Lambda}_k^{(G)} = \boldsymbol{\Lambda}_k$

$$[\boldsymbol{\Lambda}_k | \psi_{ck}^{(1)}, \cdots, \psi_{ck}^{(G)}, \boldsymbol{\mu}^{(1)}, \cdots, \boldsymbol{\mu}^{(G)}, \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{X}] \overset{D}{=} N[\mathbf{a}_k, \mathbf{A}_k], \quad (6.A14)$$

where $\mathbf{a}_k = \mathbf{A}_k (\mathbf{H}_{0gk}^{-1} \boldsymbol{\Lambda}_{0k} + \sum_{g=1}^{G} \psi_{ck}^{(g)-1} \boldsymbol{\Omega}^{(g)} \mathbf{V}_k^{*(g)})$, and $\mathbf{A}_k = (\sum_{g=1}^{G} \psi_{ck}^{(g)-1} \boldsymbol{\Omega}^{(g)} \boldsymbol{\Omega}^{(g)T} + \mathbf{H}_{0gk}^{-1})^{-1}$.

② under $\psi_{ck}^{(1)} = \cdots = \psi_{ck}^{(G)} = \psi_{ck}$,

$$[\psi_{ck}^{-1} | \boldsymbol{\Lambda}_k^{(1)}, \cdots, \boldsymbol{\Lambda}_k^{(G)}, \boldsymbol{\mu}^{(1)}, \cdots, \boldsymbol{\mu}^{(G)}, \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{X}] \overset{D}{=} Gamma(N^*/2 + \alpha_{0ck}, \beta_{ck}), \quad (6.A15)$$

where $N^* = N_1 + \cdots + N_G$,

$$\beta_{ck} = \beta_{0ck} + \sum_{g=1}^{G} [\boldsymbol{\Lambda}_k^{(g)T} (\boldsymbol{\Omega}^{(g)} \boldsymbol{\Omega}^{(g)T}) \boldsymbol{\Lambda}_k^{(g)} - 2\boldsymbol{\Lambda}_k^{(g)T} \boldsymbol{\Omega}^{(g)} \mathbf{V}_k^{*(g)} + \mathbf{V}_k^{*(g)T} \mathbf{V}_k^{*(g)}]/2.$$

③ Under $\boldsymbol{\Phi}^{(1)} = \cdots = \boldsymbol{\Phi}^{(G)} = \boldsymbol{\Phi}$,

$$[\boldsymbol{\Phi} | \boldsymbol{\Omega}_2^{(1)}, \cdots, \boldsymbol{\Omega}_2^{(G)}] \overset{D}{=} IW_{q_2}[(\sum_{g=1}^{G} \boldsymbol{\Omega}_2^{(g)} \boldsymbol{\Omega}_2^{(g)T} + \mathbf{R}_0^{-1}),\ N^* + \rho_0].$$

Conditional Distributions: Two-level Nonlinear SEM

① $$p(\mathbf{v}_{ij} | \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{Y}_{gj}, \boldsymbol{\Omega}_{1g}, \boldsymbol{\omega}_{2g}, \mathbf{X}_{gj}, \mathbf{Z}_g) \propto p(\mathbf{v}_{ij} | \boldsymbol{\theta}, \boldsymbol{\omega}_{2g}) \prod_{i=1}^{N_j} p(\mathbf{u}_{ji} | \boldsymbol{\theta}, \mathbf{v}_{ji}, \boldsymbol{\omega}_{1gj}) \quad (6.A1)$$

$$\propto \exp\left[-\frac{1}{2} \left\{ \mathbf{v}_{ij}^T (N_j \boldsymbol{\Psi}_{1g}^{-1} + \boldsymbol{\Psi}_{2}^{-1}) \mathbf{v}_{ij} - 2\mathbf{v}_{ij}^T \left[\boldsymbol{\Psi}_{1g}^{-1} \sum_{i=1}^{N_j} (\mathbf{u}_{ji} - \boldsymbol{\Lambda}_{1g} \boldsymbol{\omega}_{1gj}) + \boldsymbol{\Psi}_2^{-1} (\boldsymbol{\mu} + \boldsymbol{\Lambda}_{2g} \boldsymbol{\omega}_{2g}) \right] \right\} \right]$$

Hence, for each $\mathbf{v}_{ij}$, its conditional distribution $p(\mathbf{v}_{ij} | \cdot) \sim N[\boldsymbol{\mu}_{ij}^*, \boldsymbol{\Sigma}_{ij}^*]$, where

$$\boldsymbol{\mu}_{ij}^* = \boldsymbol{\Sigma}_{ij}^* [\boldsymbol{\Psi}_{1g}^{-1} \sum_{i=1}^{N_j} (\mathbf{u}_{ji} - \boldsymbol{\Lambda}_{1g} \boldsymbol{\omega}_{1gj}) + \boldsymbol{\Psi}_2^{-1} (\boldsymbol{\mu} + \boldsymbol{\Lambda}_{2g} \boldsymbol{\omega}_{2g})], \quad \text{and} \quad \boldsymbol{\Sigma}_{ij}^* = (N_j \boldsymbol{\Psi}_{1g}^{-1} + \boldsymbol{\Psi}_2^{-1})^{-1}.$$

② $$p(\boldsymbol{\Omega}_1 | \cdot) = \prod_{g=1}^{G} \prod_{i=1}^{N_g} p(\boldsymbol{\omega}_{1gi} | \boldsymbol{\theta}, \mathbf{v}_{gi}, \boldsymbol{\omega}_{2g}, \mathbf{u}_{gi}) \propto \prod_{g=1}^{G} \prod_{i=1}^{N_g} p(\mathbf{u}_{gi} | \boldsymbol{\theta}, \mathbf{v}_{gi}, \boldsymbol{\omega}_{1gi}) p(\boldsymbol{\eta}_{gi} | \boldsymbol{\xi}_{1gi}, \boldsymbol{\theta}) p(\boldsymbol{\xi}_{1gi} | \boldsymbol{\theta}).$$

It follows that $p(\boldsymbol{\omega}_{1gi} | \cdot)$ is proportional to

$$\exp\left[-\frac{1}{2} (\boldsymbol{\xi}_{1gi}^T \boldsymbol{\Psi}_{1g}^{-1} \boldsymbol{\xi}_{1gi} + (\mathbf{u}_{gi} - \mathbf{v}_g - \boldsymbol{\Lambda}_{1g} \boldsymbol{\omega}_{1gi})^T \boldsymbol{\Psi}_{1g}^{-1} (\mathbf{u}_{gi} - \mathbf{v}_g - \boldsymbol{\Lambda}_{1g} \boldsymbol{\omega}_{1gi}) + [\boldsymbol{\eta}_{1gi} - \boldsymbol{\Pi}_{1g} \boldsymbol{\eta}_{gi} - \boldsymbol{\Gamma}_{1g} \mathbf{F}_1(\boldsymbol{\xi}_{1gi})]^T \boldsymbol{\Psi}_{1g}^{-1} [\boldsymbol{\eta}_{1gi} - \boldsymbol{\Pi}_{1g} \boldsymbol{\eta}_{gi} - \boldsymbol{\Gamma}_{1g} \mathbf{F}_1(\boldsymbol{\xi}_{1gi})] \right] \quad (6.A2)$$

$p(\boldsymbol{\Omega}_2 | \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{Y}, \mathbf{V}, \boldsymbol{\Omega}_1, \mathbf{X}, \mathbf{Z})$: This distribution has very similar form as in $p(\boldsymbol{\Omega}_1 | \cdot)$

$$p(\boldsymbol{\alpha}, \mathbf{Y} | \cdot) = \prod_{g=1}^{G} p(\boldsymbol{\alpha}_g, \mathbf{Y}_g | \cdot) = \prod_{g=1}^{G} \prod_{k=1}^{s} p(\boldsymbol{\alpha}_{gk}, \mathbf{Y}_{gk} | \cdot),$$

$$p(\boldsymbol{\alpha}_{gk}, \mathbf{Y}_{gk} | \cdot) \propto \prod_{i=1}^{N_g} \phi\{\psi_{1ggk}^{-1/2} (y_{gik} - v_{ggk} - \boldsymbol{\Lambda}_{1ggk}^T \boldsymbol{\omega}_{1gi})\} I_{(\alpha_{gk,z_{gik}} \alpha_{gk,z_{gik}+1})}(y_{gik}),$$

④ $p(\boldsymbol{\theta} | \alpha, \gamma, \mathbf{V}, \Omega_1, \Omega_2, \mathbf{X}, \mathbf{Z})$

θ_1: $\psi_{1gk}^{-1} \overset{D}{=} Gamma[\alpha_{01gk}, \beta_{01gk}]$, $[\boldsymbol{\Lambda}_{1gk} | \psi_{1gk}] \overset{D}{=} N[\boldsymbol{\Lambda}_{01gk}, \psi_{1gk} \mathbf{H}_{01gk}]$, $k = 1, \cdots, p$,

where ψ_{1gk} is the kth diagonal element of $\boldsymbol{\Psi}_{1g}$, $\boldsymbol{\Lambda}_{1gk}^T$ is the kth row of $\boldsymbol{\Lambda}_{1g}$, and $\alpha_{01gk}, \beta_{01gk}, \boldsymbol{\Lambda}_{01gk}$, and $\mathbf{H}_{01gk}$ are given hyperparameter values. For $k \neq h$, it is assumed that $(\psi_{1gk}, \boldsymbol{\Lambda}_{1gk})$ and $(\psi_{1gh}, \boldsymbol{\Lambda}_{1gh})$ are independent. Let $\mathbf{U}_g^* = \{\mathbf{u}_{gi} - \mathbf{v}_g,\ i = 1, \cdots, N_g\}$ and $\mathbf{U}_{gk}^{*T}$ be the kth row of $\mathbf{U}_g^*$, $\boldsymbol{\Sigma}_{1gk} = (\mathbf{H}_{01gk}^{-1} + \boldsymbol{\Omega}_{1g} \boldsymbol{\Omega}_{1g}^T)^{-1}$, $\mathbf{m}_{1gk} = \boldsymbol{\Sigma}_{1gk} (\mathbf{H}_{01gk}^{-1} \boldsymbol{\Lambda}_{01gk} + \boldsymbol{\Omega}_{1g} \mathbf{U}_{gk}^*)$, $\boldsymbol{\Omega}_{1g} = (\boldsymbol{\omega}_{1g1}, \cdots, \boldsymbol{\omega}_{1gN_g})$, and $\beta_{1gk} = \beta_{01gk} + (\mathbf{U}_{gk}^{*T} \mathbf{U}_{gk}^* - \mathbf{m}_{1gk}^T \boldsymbol{\Sigma}_{1gk}^{-1} \mathbf{m}_{1gk} + \boldsymbol{\Lambda}_{01gk}^T \mathbf{H}_{01gk}^{-1} \boldsymbol{\Lambda}_{01gk})/2$, it can be shown that

$$[\psi_{1gk}^{-1} | \cdot] \overset{D}{=} Gamma\ (N_g/2 + \alpha_{01gk}, \beta_{1gk}), \quad [\boldsymbol{\Lambda}_{1gk} | \psi_{1gk}, \cdot] \overset{D}{=} N[\mathbf{m}_{1gk}, \psi_{1gk} \boldsymbol{\Sigma}_{1gk}].$$

$\psi_{1gbk}^{-1} \overset{D}{=} Gamma[\alpha_{01gbk}, \beta_{01gbk}]$, $[\boldsymbol{\Lambda}_{1gk}^* | \psi_{1gbk}] \overset{D}{=} N[\boldsymbol{\Lambda}_{01gk}^*, \psi_{1gbk} \mathbf{H}_{01gk}^*]$, $k = 1, \cdots, q_{11}$,

where α_{01gbk}, β_{01gbk}, $\boldsymbol{\Lambda}_{01gk}^*$, and $\mathbf{H}_{01gk}^*$ are given hyperparameters. Let $\boldsymbol{\Omega}_{1g}^* = \{\boldsymbol{\eta}_{1g1}, \cdots, \boldsymbol{\eta}_{1gN_g}\}$, $\boldsymbol{\Omega}_{1gk}^{*T}$ be the kth row of $\boldsymbol{\Omega}_{1g}^*$, $\boldsymbol{\Xi}_{1g} = \{\boldsymbol{\xi}_{1g1}, \cdots, \boldsymbol{\xi}_{1gN_g}\}$ and $\mathbf{F}_{1g}^* = \{\mathbf{F}_1^*(\omega_{1g1}), \cdots, \mathbf{F}_1^*(\omega_{1gN_g})\}$, in which $\mathbf{F}_1^*(\omega_{1gi}) = (\boldsymbol{\eta}_{1gi}^T, \mathbf{F}_i(\boldsymbol{\xi}_{1gi})^{*T})^T$, $i = 1, \cdots, N_g$, it can be shown that

$$[\psi_{1gbk}^{-1} | \cdot] \overset{D}{=} Gamma\ [N_g/2 + \alpha_{01gbk}, \beta_{1gbk}], \quad [\boldsymbol{\Lambda}_{1gk}^* | \psi_{1gbk}^{-1}, \cdot] \overset{D}{=} N[\mathbf{m}_{1gk}^*, \psi_{1gbk} \boldsymbol{\Sigma}_{1gk}^*], \quad (6.A6)$$

where $\boldsymbol{\Sigma}_{1gk}^* = (\mathbf{H}_{01gk}^{-1} + \mathbf{F}_{1g}^* \mathbf{F}_{1g}^{*T})^{-1}$, $\mathbf{m}_{1gk}^* = \boldsymbol{\Sigma}_{1gk}^* (\mathbf{H}_{01gk}^{-1} \boldsymbol{\Lambda}_{01gk}^* + \mathbf{F}_{1g}^* \boldsymbol{\Omega}_{1gk}^*)$, and $\beta_{1gbk} = \beta_{01gbk} + (\boldsymbol{\Omega}_{1gk}^{*T} \boldsymbol{\Omega}_{1gk}^* - \mathbf{m}_{1gk}^{*T} \boldsymbol{\Sigma}_{1gk}^{*-1} \mathbf{m}_{1gk}^* + \boldsymbol{\Lambda}_{01gk}^{*T} \mathbf{H}_{01gk}^{*-1} \boldsymbol{\Lambda}_{01gk}^*)/2$. The conditional distribution relating to $\boldsymbol{\Phi}_{1g}$ is given by

$$[\boldsymbol{\Phi}_{1g} | \boldsymbol{\Xi}_{1g}] \overset{D}{=} IW_{q_{12}}[(\boldsymbol{\Xi}_{1g} \boldsymbol{\Xi}_{1g}^T + \mathbf{R}_{01g}^{-1}), N_g + \rho_{01g}].$$

θ_2: $\mu \overset{D}{=} N[\mu_0, \Sigma_0]$, $\psi_{2k}^{-1} \overset{D}{=} Gamma[\alpha_{02k}, \beta_{02k}]$, $[\boldsymbol{\Lambda}_{2k} | \psi_{2k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{02k}, \psi_{2k} \mathbf{H}_{02k}]$.

$\boldsymbol{\Phi}_2^{-1} \overset{D}{=} W_{q_2}[\mathbf{R}_{02}, \rho_{02}]$, $\psi_{2bk}^{-1} \overset{D}{=} Gamma[\alpha_{02bk}, \beta_{02bk}]$, $[\boldsymbol{\Lambda}_{2k}^* | \psi_{2bk}] \overset{D}{=} N[\boldsymbol{\Lambda}_{02k}^*, \psi_{2bk} \mathbf{H}_{02k}^*]$,

where $\boldsymbol{\Lambda}_k^* = (\boldsymbol{\Pi}_2, \boldsymbol{\Gamma}_2)$, $\boldsymbol{\Lambda}_{02k}^{*T}$ is the vector that contains the unknown parameters in the kth row of $\boldsymbol{\Lambda}_2^*$, and ψ_{2bk} is the kth diagonal element of $\boldsymbol{\Psi}_{2b}$. As these conditional distributions are similar to those in (6.A5)–(6.A7), they are not presented here.