S&DS 410/610 : Statistical Inference

Based on lectures by Zhou Fan

Notes taken by Leqi Xu *

2021 Fall

*If you find any typos, please email them to leqi.xu@yale.edu. Many thanks to Tianyu Liu, Xuankai Wang for pointing out previous typos.

Main Topics in Statistical Inference

Unit I: Finite sample theory

Exponential family models
Natural parameters and sufficient statistics. Cumulant generating function $A(\eta)$, moments and cumulants of $T(x)$. [3]

Sufficiency, minimal sufficiency and Rao-Blackwell [1]

Decision theory and optimality of estimators
Admissibility and domination. Bayesian paradigm: average risk with respect to prior, posterior distributions, Bayes estimator and Bayes risk. Minimax paradigm: worst-case risk, Bayes risk as lower bounds, least favorable priors. [4]

High dimensions (Gaussian sequence model)
Shrinkage estimation, James-Stein estimation, Stein's Lemma and Stein's unbiased estimator for risk. Sparsity, soft thresholding and hard thresholding [2]

Hypothesis testing
Type I and type II error. Simple v.s. simple tests, likelihood ratio and Neyman-Pearson Lemma. Uniformly most power tests, average power v.s. worst-case power. Gaussian sparse model: sparse v.s. dense alternatives to $\theta = 0$. Multiple testing, false discovery rate, Bonferroni procedure and Holm's procedure. [4]

Unit II: Asymptotic theory

Basic tools
Weak Law of large Numbers. Central Limit Theorem. Slutsky's Lemma. Continuous Mapping Theorem. Delta Method. [2]

Maximum likelihood estimation and point-wise asymptotics
Asymptotic consistency, uniform convergence and covering net. Asymptotic normality, efficiency of maximum likelihood estimation and Taylor expansion of log-likelihood. [3]

Local asymptotics
Testing, contiguity and local asymptotic normality of log-likelihood. Asymptotic power under local alternatives and 3 likelihood-based tests.Limiting normal experiment, efficiency and superefficiency of estimators. [6]

Contents

Unit 1 Finite Sample Theory

Example 1. Suppose $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$. How to estimate θ?
The "obvious" answer is the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We can measure error by *loss function*

$$L(\bar{X}, \theta) = (\bar{X} - \theta)^2.$$

This is random. We can take the expected value, called the *risk*

$$R(\bar{X}, \theta) = \mathbb{E}_\theta \left[(\bar{X} - \theta)^2\right] = \frac{1}{n}.$$

Example 2. Suppose $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, $Y_1, Y_2, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$, $Z_1, Z_2, \cdots, Z_n \overset{\text{i.i.d.}}{\sim} N(\nu, 1)$. How to estimate (θ, μ, ν)?
The "obvious" answer is the sample mean

$$(\bar{X}, \bar{Y}, \bar{Z}) = \left(\frac{X_1 + X_2 + \cdots + X_n}{n}, \frac{Y_1 + Y_2 + \cdots + Y_n}{n}, \frac{Z_1 + Z_2 + \cdots + Z_n}{n}\right).$$

Total squared error risk is

$$R\left((\bar{X}, \bar{Y}, \bar{Z}), (\theta, \mu, \nu)\right) = \mathbb{E}_{\theta,\mu,\nu}\left[(\bar{X} - \theta)^2 + (\bar{Y} - \mu)^2 + (\bar{Z} - \nu)^2\right] = \frac{3}{n}.$$

Theorem 1. (Stein '56, '61) There exists an estimate $(\delta_X, \delta_Y, \delta_Z)$ which is *uniformly* better than $(\bar{X}, \bar{Y}, \bar{Z})$

$$R\left((\delta_X, \delta_Y, \delta_Z), (\theta, \mu, \nu)\right) < \frac{3}{n} \text{ for } \textit{every } (\theta, \mu, \nu).$$

One such estimator is

$$(\delta_X, \delta_Y, \delta_Z) = \left(1 - \frac{1}{n(\bar{X}^2 + \bar{Y}^2 + \bar{Z}^2)}\right)(\bar{X}, \bar{Y}, \bar{Z}).$$

The improvement is the biggest when $(\theta, \mu, \nu) = (0, 0, 0)$. This is called the "Stein's Paradox" – Efron, Morris.

Question 1. In what sense is $\bar{X}$ a "good" estimate of θ?

Answer 1.

- Gauss:
  - Among all possible θ, $\theta = \bar{X}$ maximizes the probability of data. (Maximum likelihood – developed by Fisher, 1912-1922).
  - $\bar{X}$ is unbiased: $\mathbb{E}_\theta[\bar{X}] = \theta$. It minimizes $R(\bar{X}, \theta)$ among all linear, unbiased estimators. (In fact, it's true among *all* unbiased estimators).

– Fisher, Neyman, Pearson, Wald, $\cdots$:

  ◦ Sufficiency: All "information" about θ is contained in $\bar{X}$. By Rao-Blackwell, if δ_X is another estimator, not a function of $\bar{X}$, then there exists $\delta_X^{'}$ such that

$$R(\theta, \delta_X^{'}) < R(\theta, \delta_X) \quad \forall \theta.$$

  ◦ Minimax: Using Bayesian techniques, for any estimator δ_X

$$\frac{1}{n} = \sup_{\theta \in \mathbb{R}} R(\theta, \bar{X}) \leq \sup_{\theta \in \mathbb{R}} R(\theta, \delta_X).$$

  ◦ Admissibility: Using Bayesian techniques, there does *not* exist estimator δ_X such that

$$R(\theta, \delta_X) < R(\theta, \bar{X}) \quad \forall \theta \in \mathbb{R}.$$

  (Holds in dimensions 1 or 2, not ≥ 3).

  ◦ local asymptotic normality: as $n \to \infty$, this normal model is a "good approximation" for many other parametric models.

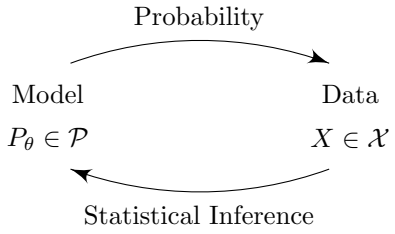
Definition 1. A *statistical model* is a family of probability distributions P over a sample space $\mathcal{X}$. Most of this class index P by a *parameter* θ belonging to a parameter space $\Omega : \mathcal{P} = \{P_\theta : \theta \in \Omega\}$.

Goal: Use data from p_θ to draw conclusions about θ.

– Point estimation: Estimate θ or a function $g(\theta)$.

– Hypothesis testing: Does θ belong to $\Omega_0 \subset \Omega$ or to $\Omega_1 \subset \Omega_1$?

– Uncertainty quantification: What is a set or interval to which θ belongs?



Not covered: Design experiments, choice of models, diagnostics for model fit.

1 Preparations

1.1 Measure and Integral

Goal: To define a common language to describe continuous and discrete models.

Definition 1.1.1 (Measure)**.** A *measure* μ on $\mathcal{X}$ assigns non-negative values (possibly 0, ∞) to subsets $A \subset \mathcal{X}$ such that

(i) if A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B). \tag{1.1.1}$$

(ii) if $\{A_i\}_{i=1}^{\infty}$ are disjoint, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i). \tag{1.1.2}$$

Example 1.1.1. For finite or countable $\mathcal{X}$, μ is called a *counting measure* on $\mathcal{X}$ if

$$\mu(A) = \begin{cases} \text{number of elements in } A & A \text{ is finite} \\ \infty & A \text{ is infinite.} \end{cases} \tag{1.1.3}$$

Example 1.1.2. Suppose $\mathcal{X} = \mathbb{R}$. Then there exists a unique measure μ on a large class of subsets (Borel σ-algebra) of $\mathcal{X}$ such that

$$\mu\left((a, b)\right) = b - a. \tag{1.1.4}$$

In this case, μ is called the *Lebesgue measure* on $\mathcal{X}$.

Remark 1.1.1.

(i) If $\mathcal{X}$ is finite or countable, we can define μ for *all subsets* of $\mathcal{X}$.

(ii) If $\mathcal{X} = \mathbb{R}$, we can only define μ on a class of subsets $\mathbb{t}$ such that

(a) $\emptyset, \mathcal{X} \in \mathbb{t}$,

(b) If $A \in \mathbb{t}$, then $A^c \in \mathbb{t}$,

(c) If $A_1, A_2, A_3, \cdots \in \mathbb{t}$, then $\sum_{i=1}^{\infty} A_i \in \mathbb{t}$. $\mathbb{t}$ is called a *σ-algebra*.

The *Borel σ-algebra* is the smallest σ-algebra containing all open (and closed) subsets of $\mathcal{X}$. These sets are the *measurable sets*.

Example 1.1.3. Suppose $\mathcal{X} = \mathbb{R}^n$, then there exists a unique measure μ on the Borel σ-algebra $\mathbb{t}$, which is the Lebesgue measure, that satisfies for any $(a_i, b_i) \subset \mathbb{t},\ i = 1, 2, \cdots, n$

$$\mu\left((a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)\right) = \prod_{i=1}^{n} (b_i - a_i). \tag{1.1.5}$$

Associated to every measure μ is a model of defining an integral on $\mathcal{X}$:

(i) $f : \mathcal{X} \to \mathbb{R}$ is *simple* if it takes only a finite number of values $\{a_1, a_2, \cdots, a_k\}$ on $\{A_1, A_2, \cdots, A_k\}$, then

$$\int f(x)\mathrm{d}\mu(x) = \sum_{i=1}^{k} a_i \cdot \mu(A_i). \tag{1.1.6}$$

(ii) If $f_1, f_2, \cdots$ are simple, non-negative functions and $f_i(x) \nearrow f(x)$ for any $x \in \mathcal{X}$, then

$$\int f(x)\mathrm{d}\mu(x) = \lim_{i\to\infty} \int f_i(x)\mathrm{d}\mu(x) \quad (\text{possibly } \infty). \tag{1.1.7}$$

(iii) For arbitary $f : \mathcal{X} \to \mathbb{R}$, let $f^+(x) = \max(f(x), 0), f^-(x) = -\min(f(x), 0)$ (Check: $f(x) = f^+(x) - f^-(x)$).If

$$\int f^+(x)\mathrm{d}\mu(x) < \infty, \ \int f^-(x)\mathrm{d}\mu(x) < \infty \iff \int |f(x)|\,\mathrm{d}\mu(x) < \infty,$$

then f is *integrable*, and

$$\int f(x)\mathrm{d}\mu(x) = \int f^+(x)\mathrm{d}\mu(x) - \int f^-(x)\mathrm{d}\mu(x). \tag{1.1.8}$$

Remark 1.1.2. For arbitary functions f, where f^+, f^- can be described by (iii) in the above example is called *measureable functions*.

Example 1.1.4. If $\mathcal{X}$ is finite or countable, μ is the counting measure, then

$$\int f(x)\mathrm{d}\mu(x) = \sum_{x\in\mathcal{X}} f(x). \tag{1.1.9}$$

And f is integrable $\iff \sum\limits_{x\in\mathcal{X}} |f(x)| < \infty$.

Example 1.1.5. Suppose $\mathcal{X} = \mathbb{R}$ and μ is the Lebesgue measure, then for "nice" functions, this is the usual integral for calculations

$$\int f(x)\mathrm{d}\mu(x) = \int f(x)\mathrm{d}x. \tag{1.1.10}$$

Similarly for the integral in $\mathbb{R}^n$.

Example 1.1.6. Let $(\mathcal{X}, \mu)$ be any space, $A \subset \mathcal{X}$ be any (measurable) set. Let $\mathbb{1}_A = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$, then $\int \mathbb{1}_A(x)\mathrm{d}\mu(x) = \mu(A)$. Write

$$\int_A f(x)\mathrm{d}\mu(x) = \int \mathbb{1}_A \cdot f(x)\mathrm{d}\mu(x). \tag{1.1.11}$$

And if $\mathcal{X}$ is countable, then $\int_A f(x)\mathrm{d}\mu(x) = \sum\limits_{x\in A} f(x)$.

Proposition 1.1.1. Basic properties of integral are the followings:

(i) Linearity:

$$\int c \cdot f(X) \mathrm{d}\mu(x) = c \cdot \int f(x) \mathrm{d}\mu(x), \tag{1.1.12}$$

$$\int (f(x) + g(x)) \, \mathrm{d}\mu(x) = \int f(x) \mathrm{d}\mu(x) + \int g(x) \mathrm{d}\mu(x)). \tag{1.1.13}$$

(ii) If $f \geq 0$, then

$$\int f(x) \mathrm{d}\mu \geq 0. \tag{1.1.14}$$

If $f(x) \geq g(x)$, then

$$\int f(x) \mathrm{d}\mu(x) \geq \int g(x) \mathrm{d}\mu(x). \tag{1.1.15}$$

(iii) Suppose $N \subset \mathcal{X}$, $\mu(N) = 0$ and $f(x) = g(x)$ for any $x \notin N$, then we say $f = g$ is *almost everywhere (a.e.)*. Then

$$\int f(x) \mathrm{d}\mu(x) = \int g(x) \mathrm{d}\mu(x). \tag{1.1.16}$$

Proposition 1.1.2. Less basic properties of integral are the followings: Suppose $f_1, f_2, \cdots : \mathcal{X} \to \mathbb{R}$ satisfy $f_i(x) \to f(x)$ for any $x \in \mathcal{X}$. The following is not always true

$$\int f_i(x) \mathrm{d}\mu = \lim_{i \to \infty} \int f_i(x) \mathrm{d}\mu(x). \tag{1.1.17}$$

Example 1.1.7. Suppose $\mathcal{X}$ is $\mathbb{R}$ and μ is the Lebesgue measure. $f_i(x) = \begin{cases} i & x \in (0, \frac{1}{i}) \\ 0 & \text{otherwise} \end{cases}$, Then $\int f_i(x) \mathrm{d}\mu(x) = 1$ for any i, But $f_i(x) \to f(x) \equiv 0$ for any $x \in \mathbb{R}$ and $\int f(x) d\mu(x) = 0$, so

$$\int \lim_{i \to \infty} f_i(x) \mathrm{d}\mu(x) \neq \lim_{i \to \infty} \int f_i(x) \mathrm{d}\mu(x). \tag{1.1.18}$$

Lemma 1.1.1 (Fatou's lemma)**.** If $f_1, f_2, \cdots \geq 0$ then

$$\lim_{i \to \infty} \int f_i(x) \mathrm{d}\mu(x) \geq \int (\lim_{i \to \infty} f_i(x)) \mathrm{d}\mu(x). \tag{1.1.19}$$

Theorem 1.1.1 (Monotone Convergence Theorem)**.** If $f_1, f_2, \cdots \geq 0$ and $f_i(x) \nearrow f(x)$ for any $x \in \mathcal{X}$, then

$$\lim_{i \to \infty} \int f_i(x) \mathrm{d}\mu(x) = \int f(x) \mathrm{d}\mu(x) \quad \text{(possibly } \infty\text{)}. \tag{1.1.20}$$

Theorem 1.1.2 (Dominated Convergence Theorem)**.** If $f_i(x) \to f(x)$ for any $x \in \mathcal{X}$ and $\exists g : \mathcal{X} \to \mathbb{R}$ such that $|f_i(x)| \leq g(x)$ for any i, x and $\int g(x) \mathrm{d}\mu(x) < \infty$, then

$$\lim_{i \to \infty} \int f_i(x) \mathrm{d}\mu(x) = \int f(x) \mathrm{d}\mu(x). \tag{1.1.21}$$

1.2 Probability and Densities

Example 1.2.1. Let X be a random variable taking value in $\mathcal{X}$. The fraction $P^x(A) = \mathbb{P}[x \in A]$ is a measure, called the *distribution* or *law* of x.

Definition 1.2.1 (Probability measure)**.** A *probability measure* μ on $\mathcal{X}$ is a measure such that $\mu(\mathcal{X}) = 1$.

Definition 1.2.2 (Density function)**.** For any space $(\mathcal{X}, \mu)$ and function $p : \mathcal{X} \to [0, \infty), \upsilon(A) = \int_A p(x)\mathrm{d}\mu(x) = \int p(x)\mathbb{1}_A(x)\mathrm{d}\mu(x)$ defines a new measure on $\mathcal{X}$. $p(x)$ is the *density function* of υ with respect to μ.

Example 1.2.2. If $\mathcal{X}$ is countable, and X is a "discrete" random variable on $\mathcal{X}$ with law P^x. The *probability mass function* $p(x) = \mathbb{P}[X = x]$ is the density of P^x with respect to counting measure μ and

$$P^x(A) = \mathbb{P}[X \in A] = \sum_{x \in A} \mathbb{P}[X = x] = \int_A p(x)\mathrm{d}\mu(x) \tag{1.2.1}$$

Example 1.2.3. If $\mathcal{X}$ is $\mathbb{R}$ or $\mathbb{R}^n$, X is a "continuous" random variable on $\mathcal{X}$ with law $\mathbb{P}^x$. The *probability density function* is the density of P^x with respect to Lebesgue measure μ and

$$P^x(A) = \mathbb{P}[X \in A] = \int_A p(x)\mathrm{d}x = \int_A p(x)\mathrm{d}\mu(x). \tag{1.2.2}$$

Proposition 1.2.1. If υ has density $p(x)$ with respect to μ, then

$$\int g(x)\mathrm{d}\upsilon(x) = \int g(x)p(x)\mathrm{d}\upsilon(x). \tag{1.2.3}$$

Proof. This is the proof where g is simple. Let $g(x) = \sum_{i=1}^{k} a_i \mathbb{1}_{A_i} \upsilon(A_i)(x)$, then

$$\begin{aligned}
\int g(x)\mathrm{d}\upsilon(x) &= \sum_{i=1}^{k} a_i \int \mathbb{1}_{A_i}(x)\mathrm{d}\upsilon(x) \\
&= \sum_{i=1}^{k} a_i \upsilon(A_i) \\
&= \sum_{i=1}^{k} a_i \int_{A_i} p(x)\mathrm{d}\mu(x) \\
&= \int \sum_{i=1}^{k} a_i \mathbb{1}_{A_i}(x) p(x)\mathrm{d}\mu(x) \\
&= \int g(x)p(x)\mathrm{d}\mu(x).
\end{aligned} \tag{1.2.4}$$

□

Definition 1.2.3 (Expection)**.** If X is a random variable on $\mathcal{X}$, *expectation* with respect to X is integration with respect to P^x.

Example 1.2.4.

(i) If X is discrete, then

$$\mathbb{E}[T(X)] = \int T(x)\mathrm{d}P^x(x) = \int T(x)p(x)\mathrm{d}\mu(x) = \sum_{x\in\mathcal{X}} T(x)p(x). \quad (1.2.5)$$

(ii) If X is continuous, then

$$\mathbb{E}[T(X)] = \int T(x)\mathrm{d}P^x(x) = \int T(x)p(x)\mathrm{d}x. \quad (1.2.6)$$

1.3 Exponential Family Models

Recall: A statistical model $\mathcal{P} = \{P_\theta : \theta \in \omega\}$ is a family of probability distributions on the sample space $\mathcal{X}$.

Definition 1.3.1 (Domination)**.** $\mathcal{P}$ is *dominated* by a common measure μ if each P_θ admits a density with respect to μ.

Note 1.3.1. For us, μ is usually the counting or Lebesgue measure.

Definition 1.3.2 (Identifiability)**.** θ is *identifiable* if $\theta_1 \neq \theta_2$ implies $P_{\theta_1} \neq P_{\theta_2}$.

Definition 1.3.3 (Exponential family model)**.** $\{P_\theta : \theta \in \Omega\}$ are an *exponential family model* if they have densities of the form

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta)T_i(x) - B(\theta)\right)h(x), \quad (1.3.1)$$

with respect to a common measure μ. $\eta_1, \eta_2, \cdots, \eta_s$ are called *natural parameters*, $T_1, T_2, \cdots, T_s$ are called the *sufficient statistics*. $e^{B(\theta)}$ is the normalizing constant:

$$e^{B(\theta)} = \int \exp\left(\sum_{i=1}^{s} \eta_i(\theta)T_i(x)\right)h(x)\mathrm{d}\mu(x). \quad (1.3.2)$$

Example 1.3.1. For Binomial(n, θ) model, where n is fixed and known, densities:

$$\begin{aligned} p_\theta(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\exp\left(x\log(\theta) + (n-x)\log(1-\theta)\right) \\ &= \binom{n}{x}\exp\left(x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right). \end{aligned}$$

Here, we can have different representation (1.3.1) for the density. For example,

(i) $h(x) = \binom{n}{x}, T_1(x) = x, \eta_1(\theta) = \log(\theta), T_2(x) = n - x, \eta_2(\theta) = \log(1-\theta), B(\theta) = 0.$

(ii) $h(x) = \binom{n}{x}, T_1(x) = x, \eta_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right), B(\theta) = -n\log(1-\theta).$

Remark 1.3.1.

(i) The representation (1.3.1) is not unique.

(ii) (1.3.1) is *minimal* if neither $\eta_1, \eta_2, \cdots, \eta_s$ nor $T_1, T_2, \cdots, T_s$ obey any linear relations. Otherwise, the number of terms s in (1.3.1) can be reduced.

Example 1.3.2. For $N(\mu, \sigma^2)$ model, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, densities:

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right). \end{aligned}$$

Here, $\eta_1(\theta) = -\frac{1}{2\sigma^2}, T_1(x) = x^2, \eta_2(\theta) = \frac{\mu}{\sigma^2}, T_2(x) = x, b(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$.

Oftentimes, it is convenient to reparametrize by $\eta_1, \eta_2 \cdots, \eta_s$:

$$p(x|\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x). \tag{1.3.3}$$

This is called *canonical form*. $A(\eta)$ is the *log-partition function*(a.k.a., *cumulant generating function*, *free energy*). $e^{A(\eta)}$ is the *partition function*.

$$e^{A(\eta)} = \int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) \mathrm{d}\mu(x). \tag{1.3.4}$$

Example 1.3.3. According to (1.3.1), $Binomial(n, \theta)$ model has densitites:

$$p_\theta(x) = \binom{n}{x} \exp\left(x \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right),$$

and we reparametrize it by $\eta = \log\left(\frac{\theta}{1-\theta}\right)$, which is the *logit* transform. Then $\frac{\theta}{1-\theta} = e^\eta, \theta = \frac{e^\eta}{1+e^\eta}, \log(1-\theta) = \log\left(\frac{1}{1+e^\eta}\right) = -\log(1+e^\eta)$, so

$$p(x|\eta) = \exp(\eta x - n\log(1+e^\eta))\binom{n}{x}. \tag{1.3.5}$$

Here, $T_1(x) = x, A(\eta) = n\log(1+e^\eta), h(x) = \binom{n}{x}$.
This is well-defined for all $\eta \in \mathbb{R}$, so $\mathbb{R}$ is called the natural parameter space.

Definition 1.3.4 (Natural parameter space)**.** The *natural parameter space* is

$$\Xi = \left\{\eta \in \mathbb{R}^s : \int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) \mathrm{d}\mu(x) < \infty\right\}. \tag{1.3.6}$$

This is the largest set of natural parameters under which the model is well-defined.

Example 1.3.4. According to (1.3.2), $N(\mu, \sigma^2)$ model, $\theta = (\mu, \sigma^2)$ has densities:

$$p_\theta(x) = \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right).$$

Set $\eta_1 = -\frac{1}{2\sigma^2}, \eta_2 = \frac{\mu}{\sigma^2}$, then $\sigma^2 = -\frac{1}{2\eta_1}, \mu = -\frac{\eta_2}{2\eta_1}$, so

$$p(x|\eta) = \exp\left(\eta_1 x^2 + \eta_2 x + \frac{{\eta_2}^2}{2\eta_1} - \frac{1}{2}\log\left(-\frac{\pi}{\eta_1}\right)\right). \tag{1.3.7}$$

Note: $\int e^{\eta_1 x^2 + \eta_2 x} \mathrm{d}x < \infty \iff \eta_1 < 0$.
So $\Xi = (-\infty, 0) \times \mathbb{R}$.

Example 1.3.5. Multinomial$(n; p_0, p_1, \cdots, p_s)$ model on $s+1$ possible outcomes. Here, $\mathcal{X} = \left\{(x_0, x_1, \cdots, x_s) : 0 \le x_0 \le x_1, \cdots, x_s \le n, \sum_{i=0}^{s} x_i = n\right\}, \theta = (p_0, p_1, \cdots, p_s)$, then

$$\begin{aligned} p_\theta(x) &= \binom{n}{x_0, x_1, \cdots, x_s} p_0^{x_0} p_1^{x_1} \cdots p_s^{x_s} \\ &= \binom{n}{x_0, x_1, \cdots, x_s} \exp\left(\sum_{i=1}^{s} x_i \log(p_i) + (n - x_1 - x_2 - \cdots - x_s \log(p_0))\right) \\ &= \binom{n}{x_0, x_1, \cdots, x_s} \exp\left(\sum_{i=1}^{s} x_i \log\left(\frac{p_i}{p_0}\right) + n\log(p_0)\right) \end{aligned} \tag{1.3.8}$$

Set $\eta_i = \log\left(\frac{p_i}{p_0}\right), i = 1, 2, \cdots, s$, then $e^{\eta_i} = \frac{p_i}{p_0}$. We have $p_0 + p_1 + \cdots + p_s = p_0(1 + e^{\eta_1} + e^{\eta_2} + \cdots + e^{\eta_s}) = 1$, so $p_0 = \frac{1}{1 + e^{\eta_1} + e^{\eta_2} + \cdots + e^{\eta_s}}$, then

$$p(x|\eta) = \binom{n}{x_0, x_1, \cdots, x_s} \exp\left(\sum_{i=1}^{s} \eta_i x_i - n\log(1 + e^{\eta_1} + e^{\eta_2} + \cdots + e^{\eta_s})\right). \tag{1.3.9}$$

So $\Xi = \mathbb{R}^s$.

Example 1.3.6. For Gamma(a, b) model, $\mathcal{X} = (0, \infty)$, then

$$\begin{aligned} p_\theta(x) &= \frac{1}{\Gamma(a) b^a} x^{a-1} e^{-\frac{x}{b}} \\ &= \exp\left(a\log(x) - \frac{1}{b}x - \log(\Gamma(a)) - a\log(b)\right)\frac{1}{x}. \end{aligned} \tag{1.3.10}$$

Set $\eta_1 = a, \eta_2 = -\frac{1}{b}$, then

$$p(x|\eta) = \exp(\eta_1 \log(x) + \eta_2 x - \log(\Gamma(\eta_1)) + \eta_1 \log(-\eta_2))\frac{1}{x}. \tag{1.3.11}$$

Note: $\int_0^\infty \exp(\eta_1 \log(x) + \eta_2 x)\frac{1}{x} dx = \int_0^\infty x^{\eta_1 - 1} e^{\eta_2 x} dx < \infty \iff \eta_2 < 0, \eta_1 > 0$.
So $\Xi = (0, \infty) \times (\infty, 0)$.

Note. There are three important inequalities that we can use in the proof:

– Cauchy-Schwarz inequality: Let μ be the probability measure, $f, g : R \to R$:

$$\int f(x)g(x)d\mu(x) \le \left[\int f(x)^2 d\mu(x)\right]^{1/2} \left[\int g(x)^2 d\mu(x)\right]^{1/2}.$$
$$\mathbb{E}[f(x)g(x)] \le \sqrt{\mathbb{E}\left(f(x)^2\right)}\sqrt{\mathbb{E}\left(g(x)^2\right)}.$$

– Hölder's inequality: for any measure υ, any non-negative functions $f, g : \mathcal{X} \to [0, \infty)$, and any $\alpha, \beta > 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, then

$$\int f(x)g(x)\mathrm{d}\upsilon(x) \leq \left(\int f(x)^{\alpha}\mathrm{d}\upsilon\right)^{\frac{1}{\alpha}} \left(\int g(x)^{\beta}\mathrm{d}\upsilon\right)^{\frac{1}{\beta}}.$$

– Jensen's inequality: Let μ be the probability measure, for a random variable X and a convex function $f : R \to R$:

$$f\left(\int x d\mu(x)\right) \leq \int f(x)d\mu(x).$$
$$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}[f\left(X\right)].$$

Proposition 1.3.1. The natural parameter space Ξ is always convex.

Proof. Recall: Ξ is convex if

$$x, y \in \Xi \implies \lambda x + (1-\lambda)y \in \Xi \quad \forall \lambda \in (0,1).$$

Let $\eta = (\eta_1, \eta_2, \cdots, \eta_s) \in \Xi,\ \eta' = (\eta_1', \eta_2', \cdots, \eta_s') \in \Xi$. Fix $\lambda \in (0,1)$.

By Hölder's inequality: For any measure υ, any non-negative functions $f, g : \mathcal{X} \to [0, \infty)$, and any $\alpha, \beta > 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, then

$$\int f(x)g(x)\mathrm{d}\upsilon(x) \leq \left(\int f(x)^{\alpha}\mathrm{d}\upsilon\right)^{\frac{1}{\alpha}} \left(\int g(x)^{\beta}\mathrm{d}\upsilon\right)^{\frac{1}{\beta}}. \tag{1.3.12}$$

Apply (1.3.12) to $\upsilon(A) = \int_A h(x)\mathrm{d}\mu(x),\ f(x) = \exp\left(\sum_{i=1}^{s} \lambda\eta_i T_i(x)\right),\ g(x) = \exp\left(\sum_{i=1}^{s}(1-\lambda)\eta_i' T_i(x)\right),\ \alpha = \frac{1}{\lambda},\ \beta = \frac{1}{1-\lambda}$, then:

$$\begin{aligned}
&\int \exp\left(\sum_{i=1}^{s}(\lambda\eta_i + (1-\lambda)\eta_i')T_i(x)\right) h(x)\mathrm{d}\mu(x) \\
&= \int f(x)g(x)\mathrm{d}\upsilon(x) \\
&\leq \left(\int f(x)^{\alpha}\mathrm{d}\upsilon\right)^{\frac{1}{\alpha}} \left(\int g(x)^{\beta}\mathrm{d}\upsilon\right)^{\frac{1}{\beta}} \\
&= \left(\int \exp\left(\sum_{i=1}^{s}\eta_i T_i(x)\right) h(x)\mathrm{d}\mu(x)\right)^{\lambda} \left(\int \exp\left(\sum_{i=1}^{s}\eta_i' T_i(x)\right) h(x)\mathrm{d}\mu(x)\right)^{1-\lambda} \\
&< \infty.
\end{aligned} \tag{1.3.13}$$

So $\lambda\eta + (1-\lambda)\eta' \in \Xi$. □

Remark 1.3.2. A specific problem might restrict the parameter space to be a smaller (possibily non-convex) subset of Ξ.

Example 1.3.7. For the normal model that we restrcit mean = st.dev. $N(\theta, \theta^2)$ $\theta > 0$. It has densities:

$$p_\theta(x) = \exp\left(-\frac{1}{2\theta^2}x^2 + \frac{1}{\theta}x - \frac{1}{2} - \frac{1}{2}\log(2\pi\theta^2)\right).$$

This representation is minimal since there are no linear relations.
Here, $(\eta_1, \eta_2) = (-\frac{1}{2\theta^2}, \frac{1}{\theta})$. Parameter space: $(\eta_1, \eta_2) \in \Xi : \eta_2 > 0. \eta_1 = -\frac{1}{2}{\eta_2}^2$.
These examples are sometimes called "curved" exponential families.

Proposition 1.3.2. $A(\eta)$ is convex.

Proof. Recall: $A(\eta)$ is convex if $\eta, \eta' \in \Xi \implies \lambda\eta + (1-\lambda)\eta' \in \Xi \quad \forall \lambda \in (0,1)$.

$$\begin{aligned}
&A(\lambda\eta + (1-\lambda)\eta') \\
&= \log\left[\int \exp\left(\sum_{i=1}^{s}(\lambda\eta_i + (1-\lambda)\eta_i')T_i(x)\right)h(x)\mathrm{d}\mu(x)\right] \\
&\le \log\left[\left(\int \exp\left(\sum_{i=1}^{s}\eta_i T_i(x)\right)h(x)\mathrm{d}\mu(x)\right)^{\lambda}\left(\int \exp\left(\sum_{i=1}^{s}\eta_i' T_i(x)\right)h(x)\mathrm{d}\mu(x)\right)^{1-\lambda}\right] \\
&= \lambda A(\eta) + (1-\lambda)A(\eta').
\end{aligned} \tag{1.3.14}$$

□

Question 1.3.1. Why are exponential families useful?

Answer 1.3.1.

(i) Defining generalized linear models.

Example 1.3.8 (Logistic regression)**.** Let $X_i \sim Bernoulli(p_i), i = 1, 2, \cdots n$. The natural parameters are $\eta_i = \log(\frac{p_i}{1-p_i}), i = 1, 2, \cdots n$. Model these by a linear model, $\eta_i = \alpha + \beta z_i$ for observed *covariates* $z_1, z_2, \cdots, z_n$.
More generally, for many 1-parameter family $X_i \sim p(x|\eta_i)$, model $\eta_i = \alpha + \beta z_i$. This is called using the *canonical link function*. The advantages are concave log-likelihood and finite number of sufficient statistics.

(ii) In Bayesian inference, we place a *prior* $p(\eta)$ on η and we use *posterior* $p(\eta|x)$. Certain forms for $p(\eta)$ are called conjugate priors, which lead to simple forms for $p(\eta|x)$ in the same family as $p(\eta)$. For η as the canonical parameter in exponential families, there is always a conjugate prior.

(iii) Useful way of extending simpler statistical models.

Example 1.3.9 (Random graphs)**.** Let $\mathcal{X} = \{$graphs on n vertices, represented by $x_1, x_2, \cdots, x_{\binom{n}{s}} \in \{0,1\}\}$. The simple model is Erdos-Renyi graph where edges are Bernoulli(θ). Set $\eta = \log(\frac{\theta}{1-\theta})$,

$$\begin{aligned}
p(x|\eta) &= \prod_{i=1}^{\binom{n}{2}} \exp(x_i\eta - \log(1+e^\eta)) \\
&= \exp\left(\eta \cdot (\text{number of edges}) - \binom{n}{2}\log(1+e^\eta)\right).
\end{aligned} \tag{1.3.15}$$

In reality, we may want to capture more features such as the number of triangles or 4-cliques in the network. Let $T_1(x) =$ number of edges, $T_2(x) =$ number of triangles, $T_3(x) =$ number of 4-cliques. Consider the model

$$p(x|\eta) = \exp(\eta_1 T_1(x) + \eta_2 T_2(x) + \eta_3 T_3(x) - A(\eta)). \tag{1.3.16}$$

This model contains Erdos-Renyi as special case ($\eta_2 = \eta_3 = 0$). Varients η_2, η_3 encourage or discourage the number of triangles or 4-cliques. This is called an *exponential random graph model* (ERGM).
Note: $A(\eta)$ typically lacks a closed form expression, and is hard to compute.

$$A(\eta) = \log\left(\sum_{i\in\{0,1\}^{\binom{n}{2}}} \exp(\eta_1 T_1(x) + \eta_2 T_2(x) + \eta_3 T_3(x))\right) \tag{1.3.17}$$

(iv) Minimize Kullback–Leibler divergence. Consider a "simple" distribution P for the data with density p. And we want to find the closest distribution Q such that $E_Q[T_i(x)] = t_i \quad \forall i = 1, 2, \cdots, s$ for some fixed values $t_1, t_2, \cdots t_s$ and certain statistics $T_1, T_2, \cdots T_s$. Let

$$D_{KL}(Q||P) = \int q(x) \log\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x), \tag{1.3.18}$$

the Kullback–Leibler divergence. Our goal is to find Q which minimizes $D_{KL}(Q||P)$ such that $E_Q[T_i(x)] = t_i \quad \forall i = 1, 2, \cdots, s$.

Claim 1.3.1. Under mild assumptions, Q belongs to exponential family.

Proof. Proof sketch: Introduce Lagrange multipliers.
Goal:

$$\begin{aligned}
&\min_{Q: E_Q[T_i(x)]=t_i \ \forall i} \{D_{KL}(Q||P)\} \\
&= \min_{Q} \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ D_{KL}(Q||P) + \sum_{i=1}^{s} \eta_i (t_i - E_Q[T_i(x)]) \right\}
\end{aligned}$$

If Q satisfies mild convex density conditions

$$\begin{aligned}
\text{Goal} &= \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \min_{Q} \left\{ D_{KL}(Q||P) + \sum_{i=1}^{s} \eta_i (t_i - E_Q[T_i(x)]) \right\} \\
&= \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ \sum_{i=1}^{s} \eta_i t_i + \min_{Q} \left\{ D_{KL}(Q||P) - \sum_{i=1}^{s} \eta_i E_Q[T_i(x)] \right\} \right\} \\
&= \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ \sum_{i=1}^{s} \eta_i t_i + \min_{Q} \int \left\{ \log(\frac{q(x)}{p(x)}) - \sum_{i=1}^{s} \eta_i T_i(x) \right\} q(x) \mathrm{d}\mu(x) \right\}.
\end{aligned}$$

Let $p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) p(x)$.

Then $\log(p_\eta(x)) = \sum_{i=1}^{s} \eta_i T_i(x) + \log(p(x)) - A(\eta))$.

$$\begin{aligned}
&\max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ \sum_{i=1}^{s} \eta_i t_i + \min_Q \int \left\{ \log(\frac{q(x)}{p(x)}) - \sum_{i=1}^{s} \eta_i T_i(x) \right\} q(x) \mathrm{d}\mu(x) \right\} \\
&= \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ \sum_{i=1}^{s} \eta_i t_i + \min_Q \int \left\{ \log(q(x)) - \log(p_\eta(x)) - A(\eta) \right\} q(x) \mathrm{d}\mu(x) \right\} \\
&= \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ \sum_{i=1}^{s} \eta_i t_i - A(\eta) + \min_Q \int \left\{ \log \frac{q(x)}{p_\eta(x)} \right\} q(x) \mathrm{d}\mu(x) \right\} \\
&= \max_{\eta_1,\eta_2,\cdots,\eta_s\in\mathbb{R}} \left\{ \sum_{i=1}^{s} \eta_i t_i - A(\eta) + \min_Q D_{KL}(Q||p_\eta) \right\}.
\end{aligned} \tag{1.3.19}$$

Since $D_{KL}(Q||p_\eta) > 0$, we need $Q = P_\eta$ and η maxmizes $\sum_{i=1}^{s} \eta_i t_i - A(\eta)$. □

(v) If $x_1, x_2, \cdots, x_n \overset{\text{i.i.d.}}{\sim} P_\eta$, where $P_\eta(x_i)$ are exponential family models $p_\eta(x_i) = \exp(\sum_{j=1}^{s} \eta_j T_j(x_i) - A(\eta))h(x_i)$. then

$$\begin{aligned}
p(x_1, x_2, \cdots, x_n|\eta) &= \prod_{i=1}^{n} \exp\left(\sum_{j=1}^{s} \eta_j T_j(x_i) - A(\eta) \right) h(x_i) \\
&= \exp\left\{ \sum_{j=1}^{s} \eta_j \left(\sum_{i=1}^{n} T_j(x_i) - nA(\eta) \right) \right\} \prod_{i=1}^{n} h(x_i).
\end{aligned} \tag{1.3.20}$$

This is still an exponential family model on the space $\mathcal{X}^n$ with same number of sufficient statistics which are now $\sum_{i=1}^{n} T_j(x_i)$ for $j = 1, 2, \cdots s$.

1.4 Moments and Cumulants

Recall: An exponential family model $\mathcal{P} = \mathbb{P}_\theta : \theta \in \Omega$ has the densities

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - b(\theta) \right) h(x)$$

with respect to a common measure μ. It's often convenient to reparametrize by $\eta_1, \eta_2, \cdots, \eta_s$;

$$p(x|\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta) \right) h(x).$$

Here, $e^{A(\eta)} = \int \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - b(\theta) \right) h(x) \mathrm{d}\mu(x)$ is the *partition function*. $A(\eta)$ is the *log-partition function* (a.k.a., *cumulant generating function*, *free energy*).

Definition 1.4.1 (Moment generating function)**.** The *moment generating function* of a real-valued random variable T is $M_T(u) = \mathbb{E}[e^{uT}]$.

Proposition 1.4.1. If $M_T(u) < \infty$ for any $u \in (-\delta, \delta)$ for some $\delta > 0$, then

(i) For all $\mu \in (-\delta, \delta)$, $k \geq 1$,

$$\frac{d^k}{du^k} M_T(u) = \mathbb{E}\left[T^k e^{uT}\right] \text{ and } \mathbb{E}\left[|T|^k e^{uT}\right] < \infty \tag{1.4.1}$$

(ii) If $\mathbb{E}\left[|T|^k\right] < \infty$ for any $k \geq 0$, then the Taylor expansion of $M_T(u)$ around $u = 0$ is

$$M_t(u) = 1 + (\mathbb{E}[T])\, u + \left(\mathbb{E}\left[T^2\right]\right) \frac{u^2}{2} + \left(\mathbb{E}\left[T^3\right]\right) \frac{u^3}{6} + \cdots \tag{1.4.2}$$

Proof. Applying (i) at $u = 0$ yields (ii). To show (i)

– For $k = 1$,

$$\begin{aligned}
\frac{d}{du} M_T(u) &= \lim_{h \to 0} \frac{M_T(u+h) - M_T(u)}{h} \\
&= \lim_{h \to 0} \frac{\mathbb{E}\left[e^{(u+h)T}\right] - \mathbb{E}\left[e^{uT}\right]}{h} \\
&= \lim_{h \to 0} \int \frac{e^{ut}(e^{ht} - 1)}{h} dP^T(t).
\end{aligned}$$

For all $t \in \mathbb{R}, |h| \leq \epsilon$, we have

$$\left|\frac{e^{ht} - 1}{h}\right| \leq \frac{e^{\epsilon|t|} - 1}{\epsilon} \leq \frac{e^{\epsilon|t|}}{\epsilon} \leq \frac{e^{\epsilon t} + e^{-\epsilon t}}{\epsilon}.$$

Let $g(t) = \frac{1}{\epsilon} e^{ut}(e^{\epsilon t} + e^{-\epsilon t})$, for fixed $\epsilon < \delta - |u|$. By assumption,

$$\mathbb{E}\left[|T|\, e^{uT}\right] \leq \mathbb{E}[g(T)] = \int g(t) \mathrm{d}P^T(t) = \frac{1}{\epsilon}\left(\mathbb{E}\left[e^{(u+\epsilon)t}\right] + \mathbb{E}\left[e^{(u-\epsilon)t}\right]\right) < \infty.$$

By Dominated Convergence Theorem,

$$\begin{aligned}
\frac{d}{du} M_T(u) &= \int \lim_{h \to 0} \frac{e^{ut}(e^{ht} - 1)}{h} dP^T(t) \\
&= \int \left(\frac{d}{du} e^{ut}\right) \mathrm{d}P^T(t) \\
&= \int t e^{ut} dP^T(t) = \mathbb{E}\left[T e^{uT}\right].
\end{aligned}$$

– For $k = 2$,

$$\frac{d^2}{du^2} M_T(u) = \lim_{h \to 0} \int \frac{t e^{ut}(e^{ht} - 1)}{h} dP^T(t).$$

Let $g(t) = \frac{1}{\epsilon} |t|\, e^{ut}(e^{\epsilon t} + e^{-\epsilon t})$, for fixed $\epsilon < \delta - |u|$. By the case $k = 1$,

$$\begin{aligned}
\mathbb{E}\left[\left|T^2\right| e^{uT}\right] \leq \mathbb{E}[g(T)] &= \int g(t) \mathrm{d}P^T(t) \\
&= \frac{1}{\epsilon}\left(\mathbb{E}\left[|T|\, e^{(u+\epsilon)t}\right] + \mathbb{E}\left[|T|\, e^{(u-\epsilon)t}\right]\right) < \infty.
\end{aligned}$$

By Dominated Convergence Theorem,

$$\frac{d^2}{du^2}M_T(u) = \int \lim_{h\to 0} \frac{te^{ut}(e^{ht}-1)}{h} \mathrm{d}P^T(t) = \mathbb{E}\left[T^2 e^{uT}\right].$$

– For $k \geq 3$, continue by induction.

□

Remark 1.4.1. In exponential family model

$$p(x|\eta) = \exp(\eta T(x) - A(\eta))h(x). \tag{1.4.3}$$

Suppose η_0 is in the interior of natural parameter space Ξ. Then udner $p(x|\eta_0)$, for $T = T(x)$,

$$\begin{aligned} M_T(u) &= \mathbb{E}\left[e^{uT(x)}\right] \\ &= \int e^{ut(x)} e^{\eta_0 t(x) - A(\eta_0)} h(x) \mathrm{d}\mu(x) \\ &= e^{-A(\eta_0)} \int e^{(u+\eta_0)t(x)} h(x) \mathrm{d}\mu(x) \\ &= \frac{e^{A(u+\eta_0)}}{e^{A(\eta_0)}} \end{aligned} \tag{1.4.4}$$

This is finite in a neighborhood of $u = 0$. Therefore, Moment generating functions and moments of $T(x)$ can be computed from $e^{A(\eta)}$ and its derivatives.

Example 1.4.1. For $N(\theta, 1)$ model, its densities:

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} = \exp\left(\theta x - \frac{\theta^2}{2}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Here, $\eta = \theta$ is the natural parameter, $T(x) = x$ is the sufficient statistic, $\Xi = \mathbb{R}$.

$$M_x(u) = \frac{e^{A(u+\theta)}}{e^{A(\theta)}} = \exp\left(\frac{(\theta+u)^2}{2} - \frac{\theta^2}{2}\right) = \exp\left(\frac{u^2}{2} + \theta u\right). \tag{1.4.5}$$

Taylor expand for $\theta = 0$ to get moments of $N(0, 1)$:

$$\begin{aligned} M_x(i) = \exp(\frac{u^2}{2}) &= 1 + \frac{u^2}{2} + (\frac{u^2}{2})^2 \cdot \frac{1}{2} + (\frac{u^2}{2})^3 \cdot \frac{1}{6} + \cdots \\ &= \sum_{k=0}^{\infty} \left(\frac{u^2}{2}\right)^k \cdot \frac{1}{k!} \\ &= \sum_{k=0}^{\infty} u^{2k} \cdot \frac{1}{(2k)!} \cdot \left(\frac{(2k)!}{2^k k!}\right) \end{aligned} \tag{1.4.6}$$

Therefore, $\mathbb{E}[X^{2k}] = \frac{(2k)!}{2^k k!}$. Here, $\frac{(2k)!}{2^k k!} = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2k-1) \equiv (2k-1)!!$

Moments	Cumulants
$\alpha_1 = \mathbb{E}[X]$	$\kappa_1 = \alpha_1 = \mathbb{E}[X]$
$\alpha_2 = \mathbb{E}[X^2]$	$\kappa_2 = \alpha_2 - \alpha_1^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
$\alpha_3 = \mathbb{E}[X^3]$	$\kappa_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3$

Definition 1.4.2 (Cumulant generating function)**.** Consider the Taylor expansion of $K_T(u) = \log(M_T(u))$ around $u = 0$,

$$K_T(u) = \kappa_1 u + \kappa_2 \frac{u^2}{2} + \kappa_3 \frac{u^3}{6} + \cdots \tag{1.4.7}$$

$K_T(u)$ is the *cumulant generating function* of T, and $\kappa_1, \kappa_2, \kappa_3, \cdots$ are the *cumulants* of T.

Remark 1.4.2. If T, S are independent, then

$$M_{T+S}(u) = \mathbb{E}[e^{u(T+S)}] = \mathbb{E}[e^{uT}]\mathbb{E}[e^{uS}] = M_T(u)M_S(u). \tag{1.4.8}$$

$$K_{T+S}(u) = \log(M_T(u)) + \log(M_S(u)) = K_T(u) + K_S(u). \tag{1.4.9}$$

Therefore, cumulants of T+S are these of T plus these of S. ($\mathbb{E}[T+S] = \mathbb{E}[T] + \mathbb{E}[S], Var[T+S] = Var[T] + Var[S] \cdots$)

Remark 1.4.3. In exponential family model

$$p(x|\eta) = \exp(\eta T(x) - A(\eta))h(x).$$

Suppose η_0 is in the interior of natural parameter space Ξ. Then udner $p(x|\eta_0)$, for $T = T(x)$,

$$K_T(u) = \log\left(\frac{e^{A(u+\eta_0)}}{e^{A(\eta_0)}}\right) = A(u+\eta_0) - A(\eta_0). \tag{1.4.10}$$

The k^{th} cumulant of $T(x)$ is $\kappa_k = \frac{d^k}{du^k} K_T(u)\Big|_{u=0} = \frac{d^k}{d\eta^k} A(\eta)\Big|_{\eta=\eta_0}$.

Same ideas hold in multi-parameter models:

Definition 1.4.3 (Moment generating function)**.** The *moment generating function* of a vector-valued random variable $T = (T_1, T_2, \cdots, T_s)$ is

$$M_T(u_1, u_2, \cdots, u_s) = \mathbb{E}\left[e^{u_1 T_1 + u_2 T_2 + \cdots + u_s T_s}\right].$$

Proposition 1.4.2. If $M_T(u) < \infty$ for all u in an open neighborhood of $O \in \mathbb{R}^s$, then T has finite *mixed moments* of all orders, and the Taylor expansion at $u = 0$

$$M_T(u) = \sum_{k_1, k_2, \cdots, k_s \geq 0} \mathbb{E}\left[T_1^{k_1} T_2^{k_2} \cdots T_s^{k_s}\right] \frac{u_1^{k_1} u_2^{k_2} \cdots u_s^{k_s}}{k_1! k_2! \cdots k_s!}. \tag{1.4.11}$$

Definition 1.4.4 (Cumulant generating function)**.** The *cumulant generating function* of $T = (T_1, T_2, \cdots, T_s)$ is $K_T(u_1, u_2, \cdots, u_s) = \log(M_T(u_1, u_2, \cdots, u_s))$.

Proposition 1.4.3. The *mixed cumulants* $\kappa_{k_1, k_2, \cdots, k_s}$ are the coefficients in the Taylor expansion of $K_T(u)$ around $u = 0$

$$\begin{aligned} K_T(u) &= \sum_{k_1, k_2, \cdots, k_s \geq 0} \kappa_{k_1, k_2, \cdots, k_s} \frac{u_1^{k_1} u_2^{k_2} \cdots u_s^{k_s}}{k_1! k_2! \cdots k_s!}, \\ \kappa_{k_1, k_2, \cdots, k_s} &= \frac{\partial^{k_1 + k_2 + \cdots + k_s}}{\partial u_1^{k_1} \partial u_2^{k_2} \cdots \partial u_s^{k_s}} K_T(u)\Big|_{u=0}. \end{aligned} \tag{1.4.12}$$

In particular, the 1^{st} order cumulants are $\frac{\partial}{\partial u_i} K_T(u)\Big|_{u=0} = \mathbb{E}[T_i]$, and 2^{nd} order cumulants are $\frac{\partial^2}{\partial u_i \partial u_j} K_T(u)\Big|_{u=0} = Cov[T_i, T_j]$.

Remark 1.4.4. In exponential family model

$$p(x|\eta) = \exp(\eta T(x) - A(\eta))h(x),$$

if η_0 is in the interior of Ξ:

$$\begin{aligned}
&M_T(u) = \mathbb{E}[e^{u_1T_1(x)+u_2T_2(x)+\cdots+u_sT_s(x)}] = \frac{e^{A(u+\eta_0)}}{e^{A(\eta_0)}} \\
&K_T(u) = \log(M_T(u)) = A(u+\eta_0) - A(\eta_0) \\
&\kappa_{k_1,k_2,\cdots,k_s} = \frac{\partial^{k_1+k_2+\cdots+k_s}}{\partial u_1^{k_1}\partial u_2^{k_2}\cdots\partial u_s^{k_s}} K_T(u)\bigg|_{u=0} = \frac{\partial^{k_1+k_2+\cdots+k_s}}{\partial \eta_1^{k_1}\partial \eta_2^{k_2}\cdots\partial \eta_s^{k_s}} A(\eta)\bigg|_{\eta=\eta_0}.
\end{aligned} \tag{1.4.13}$$

In particular, $\nabla A(\eta) = \mathbb{E}[T], \nabla^2 A(\eta) = Cov[T]$.

Example 1.4.2. For $Gamma(a,b)$ model: $\theta = (a,b)$, its densities:

$$p_\theta(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}}.$$

Here, $(\eta_1,\eta_2) = (a, -\frac{1}{b})$, $T(x) = (\log(x), x)$, $p(x|\eta) = \exp(\eta_1\log(x) + \eta_2 x - \log(\Gamma(\eta_1)) + \eta_1\log(-\eta_2)) \cdot \frac{1}{x}$. Therefore, $A(\eta) = \log(\Gamma(\eta_1)) - \eta_1\log(-\eta_2)$.

$$\begin{aligned}
&\mathbb{E}[X] = \frac{\partial}{\partial\eta_2} A(\eta) = -\frac{\eta_1}{\eta_2} = ab \\
&Var(X) = \frac{\partial^2}{\partial\eta_2^2} A(\eta) = -\frac{\eta_1}{\eta_2^2} = ab^2 \\
&Cov(X, \log(X)) = \frac{\partial^2}{\partial\eta_1\partial\eta_2} A(\eta) = -\frac{1}{\eta_2} = b.
\end{aligned} \tag{1.4.14}$$

Many papers devoted to studying properties or behavior of log-partition function $A(\eta)$ in complex models.

1.5 Sufficient Statistics

Definition 1.5.1 (Sufficient statistics)**.** A statistic $T(X)$ is *sufficient* in a model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if the conditional distribution of X given $T = t$ doesn't depend on θ, for any t.

Example 1.5.1. Suppose $X = (X_1, X_2)$ are i.i.d. Poisson(λ). Its densities:

$$p_\lambda(x) = \mathbb{P}[X_1 = x_1, X_2 = x_2] = \frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\frac{\lambda^{x_2}e^{-\lambda}}{x_2!} = \frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}.$$

Conditional on $T \equiv x_1 + x_2 = t$, the density of X_1, X_2 with respect to counting measure on $\{0,1,\cdots,t\} \times \{0,1,\cdots,t\}$ is

$$\mathbb{P}[X_1 = x_1, X_2 = x_2 | T = t] = \frac{\frac{\lambda^t e^{-2\lambda}}{x_1!(t-x_1)!}}{\sum_{y=0}^{t} \frac{\lambda^t e^{-2\lambda}}{y!(t-y)!}} = \frac{\frac{1}{x_1!(t-x_1)!}}{\sum_{y=0}^{t} \frac{1}{y!(t-y)!}}. \tag{1.5.1}$$

This doesn't depend on λ, so $T = X_1 + X_2$ is sufficient in this model.

Remark 1.5.1. Sufficiency implies that for any inference I draw about θ using X, you can draw just using $T(X)$. Our inferences will be equally accurate if we assume the model is correct! The procedure will be the followings:

(i) Given $T(X) = t$, you generate data X' from the conditional distribution of $X|T = t$. By sufficiency, you don't need to know θ to do this.

(ii) Apply whatever procedure I use on X to the data X'.

X and X' are equal in distribution, so our inferences have the same statistical properties including same risks of estimators, same expected coverages of intervals, same error rates for tests, etc.

Theorem 1.5.1 (Neyman-Fisher factorization criterion)**.** Suppose $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ is dominated by measure μ. A function $T(X)$ is sufficient for $\mathcal{P}$ if and only if their densities with respect to μ can be written as

$$p_\theta(x) = g_\theta(T(x))h(x) \tag{1.5.2}$$

for some function g_θ, h. Here, $h(x)$ doesn't depend on θ and dependence on θ is only via $T(x)$.

Proof.

– Proof in discrete case:

(i) Suppose $p_\theta(x) = g_\theta(T(x))h(x)$.

$$\mathbb{P}_\theta[X = x|T = t] = \begin{cases} \frac{g_\theta(t)h(x)}{\sum_{y:T(y)=t} g_\theta(t)h(y)} = \frac{h(x)}{\sum_{y:T(y)=t} h(y)} & \text{if } T(x) = t \\ 0 & \text{otherwise.} \end{cases} \tag{1.5.3}$$

doesn't depend on θ, so T is sufficient.

(ii) Suppose T is sufficient.
Let $g_\theta(t) = \mathcal{P}_\theta[T(X) = t], h(x) = \mathbb{P}_\theta[X = x|T(X) = T(x)]$. Here, $h(x)$ doesn't depend on θ because $T(X)$ is sufficient. Then

$$\begin{aligned} \mathbb{P}_\theta[X = x] &= \mathbb{P}_\theta[T(X) = T(x)]\mathbb{P}_\theta[X = x|T(X) = T(x)] \\ &= g_\theta(T(x))h(x). \end{aligned} \tag{1.5.4}$$

– Partial proof in continuous case:
Suppose $X \in \mathbb{R}^n, T(X) \in \mathbb{R}^s$ and there exists $S(X) \in \mathbb{R}^{n-s}$ such that $X \to (T(X), S(X))$ is differentiable, bijective, and has differentiable inverse. Let $J(x)$ be Jacobian of $x \to (T(x), S(x))$, $x(t, s)$ be the inverse function, and $J^{-1}(t, s)$ be its Jocobian.

(i) Suppose $p_\theta(x) = g_\theta(T(x))h(x)$. Then the density of $(T(X), S(X))$ under θ is $q_\theta(t, s) = g_\theta(t)h(x(t, s))\left|J^{-1}(t, s)\right|$. The conditional density of $S(X)$ given $T(X) = t$ is

$$q_\theta(s|T = t) = \frac{q_\theta(t, s)}{\int q_\theta(t, s)\mathrm{d}s} = \frac{h(x(t, s))\left|J^{-1}(t, s)\right|}{\int h(x(t, s))\left|J^{-1}(t, s)\right|\mathrm{d}s} \tag{1.5.5}$$

doesn't depend on θ, so the joint distribution of S, T (i.e., that of X condition on $T(x) = t$ doesn't depend on θ).

(ii) Suppose T is sufficient. Let $g_\theta(t)$ be density of $T(X)$, $\tilde{h}(t,s)$ be conditional density $q(s|T=t)$. Here, $\tilde{h}(t,s)$ doesn't depend on θ because $T(X)$ is sufficient. Then

$$\begin{aligned} &q_\theta(t,s) = g_\theta(t)\tilde{h}(t,s) \\ &p_\theta(x) = q_\theta(t(x),s(x))\,|J(x)| = g_\theta(t(x))\tilde{h}(t(x),s(x))\,|J(x)|\,. \end{aligned} \tag{1.5.6}$$

Here, $g_\theta(T(x)) = g_\theta(t(x)), h(x) = \tilde{h}(t(x),s(x))\,|J(x)|$.

□

Example 1.5.2. Suppose $X = (X_1, X_2) \overset{\text{i.i.d.}}{\sim} Poisson(\lambda)$. Its densities:

$$p_\lambda(x) = \mathbb{P}[X_1 = x_1, X_2 = x_2] = \frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}.$$

Here, $g_\lambda(T(x)) = \lambda^{x_1+x_2}e^{-2\lambda}, h(x) = \frac{1}{x_1!x_2!}$ So based on Neyman-Fisher factorization criterion, $T = X_1 + X_2$ is sufficient in this model.

Example 1.5.3. Suppose $X \sim F$ and F belongs to exponential families. Its densities:

$$p_\theta(x) = \exp\left(\sum_{j=1}^{s}\eta_j(\theta)T_j(x) - B(\theta)\right)h(x).$$

Here, $g_\theta(T(x)) = \exp\left(\sum_{j=1}^{s}\eta_j(\theta)T_j(x) - B(\theta)\right), h(x) = h(x)$. So based on Neyman-Fisher factorization criterion, $T = (T_1(X), T_2(X), \cdots, T_s(X))$ are sufficient in this model.

Example 1.5.4. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$.

$$p_\theta(x_1, x_2, \cdots, x_n) = \exp\left(\left(\sum_{j=1}^{s}\eta_j(\theta)\left(\sum_{i=1}^{n}T_j(x_i)\right)\right) - nB(\theta)\right)\prod_{i=1}^{n}h(x_i).$$

Here, $g_\theta(T(x)) = \exp\left(\left(\sum_{j=1}^{s}\eta_j(\theta)\left(\sum_{i=1}^{n}T_j(x_i)\right)\right) - nB(\theta)\right), h(x) = \prod_{i=1}^{n}h(x_i)$.
So by Neyman-Fisher factorization criterion, $T = \left(\sum_{i=1}^{n}T_1(X_i), \sum_{i=1}^{n}T_2(X_i), \cdots, \sum_{i=1}^{n}T_s(X_i)\right)$ are sufficient in this model.

Example 1.5.5. In logistic regression, $X_i \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta_i)$. $\eta_i \equiv \log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta z_i$ for fixed covariates $Z_1, Z_2, \cdots, Z_n$. Its densities:

$$\begin{aligned} p(x|\eta) &= \prod_{i=1}^{n}\exp\left((\alpha+\beta z_i)x_i - \log\left(1+e^{\alpha+\beta z_i}\right)\right) \\ &= \exp\left(\alpha\sum_{i=1}^{n}x_i + \beta\sum_{i=1}^{n}z_ix_i - \sum_{i=1}^{n}\log(1+e^{\alpha+\beta z_i})\right). \end{aligned}$$

Here, $g_{\alpha,\beta}(T(x)) = \exp\left(\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} z_i x_i - \sum_{i=1}^{n} \log(1 + e^{\alpha+\beta z_i})\right), h(x) = 1.$ So based on Neyman-Fisher factorization criterion, $T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Z_i X_i\right)$ are sufficient in this model.

Example 1.5.6. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Uniform(0, \theta)$. Its densities:

$$p_\theta(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}\{0 \le x_i \le \theta\} = \frac{1}{\theta^n} \mathbb{1}\{x_{(n)} \le \theta\} \prod_{i=1}^{n} \mathbb{1}\{x_i \le 0\}.$$

Here, $g_\theta(T(x)) = \frac{1}{\theta^n} \mathbb{1}\{x_{(n)} \le \theta\}, h(x) = \prod_{i=1}^{n} \mathbb{1}\{x_i \le 0\}$. So based on Neyman-Fisher factorization criterion, $T = X_{(n)} = \max(X_1, X_2, \cdots, X_n)$ is sufficient in this model.

Example 1.5.7. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p(x)$. Here, $p \equiv$ all continuous distribution on $\mathbb{R}$ and $\theta \equiv p$.

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} p\left(x_{(i)}\right),$$

where $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ are the order statistics. Here, $g_p(T(x)) = \prod_{i=1}^{n} p(x_{(i)}), h(x) = 1$. So based on Neyman-Fisher factorization criterion, $T = \left(X_{(1)}, X_{(2)}, \cdots, X_{(n)}\right)$ are sufficient in this model.

Example 1.5.8. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p(x)$. Here, $p \equiv$ all continuous distribution on $\mathbb{R}$ symmetric about $0, p(x) = p(-x)$ and $\theta \equiv p$.

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} p\left(|x|_{(i)}\right),$$

where $|x|_{(1)} \le |x|_{(2)} \le \cdots \le |x|_{(n)}$ are the order statistics of $|x_1|, |x_2|, \cdots, |x_n|$. Here, $g_p(T(x)) = \prod_{i=1}^{n} p\left(|x|_{(i)}\right), h(x) = 1$. So based on Neyman-Fisher factorization criterion, $T = (|X_1|, |X_2|, \cdots, |X_n|)$ are sufficient in this model.

Definition 1.5.2 (Ancillary)**.** A statistic $S(X)$ is *ancillary* for $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ if its distribution doesn't on θ.

In the above example, $(\text{sign}(x_1), \text{sign}(x_2), \cdots, \text{sign}(x_n))$ is ancillary.

1.6 Minimal Sufficiency

Example 1.6.1. Suppose $X \sim N(0, \sigma^2)$. Its densities:

$$p_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Here, $g_{\sigma^2}(T(x)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right), h(x) = 1$. So based on Neyman-Fisher factorization criterion, $T = X^2$ is sufficient in this model. So are $|X|, X^4, e^{X^2}, \cdots$ They are *equivalent*.

Definition 1.6.1 (Equivalent)**.** $T_1(X)$ and $T_2(X)$ are *equivalent* with respect to $\mathcal{P}$ if there are functions f, g such that $T_1(X) = f(T_2(X))$ and $T_2(X) = g(T_1(X))$ $\mathcal{P} - a.e.$ (There exists $N \in \mathcal{X}$ with $P(N) = 0$ for all $P \in \mathcal{P}$ such that these equalities hold for all $x \notin N$).

Note 1.6.1. If $T_1(X)$ is sufficient, then so is $T_2(X)$.

Example 1.6.2. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Its densities:

$$p_{\sigma^2}(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{2\sigma^2}\right).$$

Here, $g_{\sigma^2}(T(x)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x_1^2+x_2^2+\cdots+x_n^2}{2\sigma^2}\right), h(x) = 1$. So based on Neyman-Fisher factorization criterion,

$$\begin{aligned} T_1(x) &= (X_1, X_2, \cdots, X_n) \\ T_2(x) &= (|X|_{(1)}, |X|_{(2)}, \cdots, |X|_{(n)}) \\ T_3(x) &= X_1^2 + X_2^2 + \cdots + X_n^2 \end{aligned} \tag{1.6.1}$$

are all sufficient in this model. But they are not equivalent. T_3 is a function of T_2 which is a function of T_1, but not vice versa.

Definition 1.6.2 (Minimal sufficient)**.** A sufficient statistic $T(X)$ is *minimal sufficient* if for any other sufficient statistic $U(X)$, there is a function h such that $T(X) = h(U(X))$ $\mathcal{P} - a.e.$

Note 1.6.2.

(i) In Example 1.6.2, T_1, T_2 are not minimal sufficient, because they are not a function of T_3.

(ii) If T_1 is minimal sufficient and T_2 is equivalent to T_1, then T_2 is also minimal sufficient.

(iii) Minimal sufficiency $\iff$ Greatest reduction of data.

(iv) Not necessarily related to number of *statistics*. T is minimal suffcient $\implies$ so is (T, T^2).

Question 1.6.1. In Example 1.6.2, how to show T_3 is minimal sufficient?

Answer 1.6.1. Our idea is to consider a smaller model $\mathcal{P}_0 \in \mathcal{P}$.

(i) If $T(X)$ is sufficient in $\mathcal{P}$, it's sufficient in $\mathcal{P}_0$.

(ii) If $\mathcal{P}_0 - a.e.$ implies $\mathcal{P} - a.e.$ ($P_0(N) = 0\ \forall P_0 \in \mathcal{P}_0 \implies P(N) = 0\ \forall P \in \mathcal{P}$), and $T(X)$ is sufficient in $\mathcal{P}$ and minimal sufficient in $\mathcal{P}_0$, then $T(X)$ is minimal sufficient in $\mathcal{P}$.

Proof. Let $U(x)$ be sufficient in $\mathcal{P}$. Then $U(x)$ is sufficient in $\mathcal{P}_0$, so there exists h such that $T(X) = h(U(X))$ $\mathcal{P}_0 - a.e.$. So $T(X) = h(U(X))$ $\mathcal{P} - a.e.$ □

Theorem 1.6.1. Let $\mathcal{P} = \{P_0, P_1, \cdots, P_k\}$ have densities with respect to μ and same support. Then

$$T(X) = \left(\frac{p_1(X)}{p_0(X)}, \frac{p_2(X)}{p_0(X)}, \cdots, \frac{p_k(X)}{p_0(X)}\right) \tag{1.6.2}$$

is minimal sufficient.

Proof.

To show T is sufficient: Let $g_\theta(T(x)) = \begin{cases} 1 & \theta = 0 \\ T_\theta(x) & \theta = 1, 2, \cdots, k \end{cases}$, then

$$p_\theta(x) = g_\theta(T(x))p_0(x) \tag{1.6.3}$$

for any $\theta = 0, 1, \cdots, k$. T is sufficient by Neyman-Fisher factorization criterion. To show T is minimal: Let U be any sufficient statistics. By Neyman-Fisher factorization criterion, there are function $\tilde{g_\theta}, h$ such that

$$p_\theta(x) = \tilde{g_\theta}(U(x))h(x). \tag{1.6.4}$$

Then for each $\theta = 1, 2, \cdots, k$,

$$\frac{p_\theta(x)}{p_0(x)} = \frac{\tilde{g_\theta}(U(x))}{\tilde{g_0}(U(x))} \tag{1.6.5}$$

is a function of $U(x)$. So $T(X)$ is a function of $U(X)$. □

Example 1.6.3. $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$. To show $T_3(X) = X_1^2 + X_2^2 + \cdots + X_n^2$ is minimal sufficient, consider $\mathcal{P}_0 = \{P_0^n, P_1^n\}$ where P_0 where $P_0 = N(0, 1), P_1 = N(0, 2)$.

$$T(X) = \frac{p_1(X_1, X_2, \cdots, X_n)}{p_0(X_1, X_2, \cdots, X_n)} = \left(\frac{\sqrt{2\pi}}{\sqrt{4\pi}}\right)^n \exp\left(\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{4}\right) \tag{1.6.6}$$

is minimal sufficient for $\mathcal{P}_0$. Since T is equivalent to T_3, T_3 is minimal sufficient for $\mathcal{P}_0$. Then, T_3 is minimal sufficient in the original model.

Corollary 1.6.1. Consider the exponential family

$$p_\eta(x) = \exp(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta))h(x)$$

Suppose there are $s + 1$ points $\eta^{(0)}, \eta^{(1)}, \cdots, \eta^{(s)}$ that are affinely independent (i.e., $\eta^{(1)} - \eta^{(0)}, \eta^{(2)} - \eta^{(0)}, \cdots, \eta^{(s)} - \eta^{(0)}$ span $\mathbb{R}^s$). Then $(T_1(X), T_2(X),$ $\cdots, T_s(X))$ is minimal sufficient.

Proof. Consider $\mathcal{P}_0 = \{P_{\eta^{(0)}}, P_{\eta^{(1)}}, \cdots, P_{\eta^{(s)}}\}$. Then

$$U(X) = \left(\frac{p_{\eta^{(1)}}}{p_{\eta^{(0)}}}, \frac{p_{\eta^{(2)}}}{p_{\eta^{(0)}}}, \cdots, \frac{p_{\eta^{(s)}}}{p_{\eta^{(0)}}}\right) \tag{1.6.7}$$

is minimal sufficient in $\mathcal{P}_0$. We have

$$U_j(x) = \exp\left(\sum_{i=1}^{s} (\eta_i^{(j)} - \eta_i^{(0)})T_i(x) - A(\eta^{(j)}) + A(\eta^{(0)})\right), \tag{1.6.8}$$

so U is equivalent to

$$\left(\sum_{i=1}^{s}\left(\eta_i^{(1)}-\eta_i^{(0)}\right)T_i(x),\sum_{i=1}^{s}\left(\eta_i^{(2)}-\eta_i^{(0)}\right)T_i(x),\cdots,\sum_{i=1}^{s}\left(\eta_i^{(s)}-\eta_i^{(0)}\right)T_i(x)\right)$$
$$=\begin{pmatrix}\eta_1^{(1)}-\eta_1^{(0)} & \eta_2^{(1)}-\eta_2^{(0)} & \cdots & \eta_s^{(1)}-\eta_s^{(0)}\\ \eta_1^{(2)}-\eta_1^{(0)} & \eta_2^{(2)}-\eta_2^{(0)} & \cdots & \eta_s^{(2)}-\eta_s^{(0)}\\ \cdots & \cdots & \cdots & \\ \eta_1^{(s)}-\eta_1^{(0)} & \eta_2^{(s)}-\eta_2^{(0)} & \cdots & \eta_s^{(s)}-\eta_s^{(0)}\end{pmatrix}\begin{pmatrix}T_1(x)\\ T_2(x)\\ \cdots\\ T_s(x)\end{pmatrix} \tag{1.6.9}$$

Since the left matrix is invertible by assumption, U is equivalent to $T(X)=(T_1(X),T_2(X),\cdots,T_s(X))$. Then T is minimal sufficient in $\mathcal{P}_0$ and also in $\mathcal{P}$. □

Corollary 1.6.2. Suppose $X_1,X_2,\cdots,X_n \overset{i.i.d.}{\sim} p_\eta(x)$. Its densities:

$$p_\eta(x_1,x_2,\cdots,x_n)=\exp\left(\sum_{j=1}^{s}\eta_j\left(\sum_{i=1}^{n}T_j(x_i)\right)-nA(\eta)\right)\prod_{i=1}^{n}h(x_i).$$

This is still an exponential family model. And under the condition of Corollary 1.6.1, $\left(\sum_{i=1}^{n}T_1(X_i),\sum_{i=1}^{n}T_2(X_i),\cdots,\sum_{i=1}^{n}T_s(X_i)\right)$ is minimal sufficient.

Question 1.6.2. Is a reduction from n samples to a fixed number of sufficient statistcs always possible?

Answer 1.6.2. For smooth models with common support, possible if and only if it is an exponential family.
Note: See *Theory of Point Estimation* Theorem 1.6.18.

Example 1.6.4. Suppose $X_1,X_2,\cdots,X_n \overset{i.i.d.}{\sim} Logistic(\theta,1)$. Its densities:

$$p_\theta(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}\frac{e^{-(x_i-\theta)}}{(1+e^{-(x_i-\theta)})^2}.$$

We want to prove order statistics $T(X)=\{X_{(1)},X_{(2)},\cdots,X_{(n)}\}$ are minimal sufficient in this model.

Proof. Consider $\theta_0=0,\theta_1,\theta_2,\cdots,\theta_n$ and $\mathcal{P}_0=\{P_{\theta_0},P_{\theta_1},\cdots,P_{\theta_n}\}$. Let $x=(x_1,x_2,\cdots,x_n)$. Suppose $U(X)=(U_1(X),U_2(X),\cdots,U_n(X))$ is minimal sufficient in $\mathcal{P}_0$, where

$$U_j(x)=\prod_{i=1}^{n}\frac{\frac{e^{-(x_i-\theta_j)}}{(1+e^{-(x_i-\theta_j)})^2}}{\frac{e^{-x_i}}{(1+e^{-x_i})^2}}=e^{n\theta_j}\prod_{i=1}^{n}\left(\frac{1+e^{-x_i}}{1+e^{-(x_i-\theta_j)}}\right)^2. \tag{1.6.10}$$

Suppose $U(x)=U(y)$. Then $\frac{1}{\sqrt{U_j(x)}}=\frac{1}{\sqrt{U_j(y)}}$ for any $j=1,2,\cdots,n$. So

$$\prod_{i=1}^{n}\frac{1+\xi e^{-x_i}}{1+e^{-x_i}}=\prod_{i=1}^{n}\frac{1+\xi e^{-y_i}}{1+e^{-y_i}} \tag{1.6.11}$$

for $\xi = 1, e^{\theta_1}, \cdots, e^{\theta_n}$. These are two polynomials in ξ of degree n, which agree at $n+1$ points, so they agree at *all* ξ.
Set $\xi = 0$, then

$$\begin{aligned}
&\prod_{i=1}^{n}(1+e^{-x_i}) = \prod_{i=1}^{n}(1+e^{-y_i}) \\
&\prod_{i=1}^{n}(1+\xi e^{-x_i}) = \prod_{i=1}^{n}(1+\xi e^{-y_i}) \text{ for all } \xi \in \mathbb{R} \qquad (1.6.12) \\
&\prod_{i=1}^{n}(\eta+e^{-x_i}) = \prod_{i=1}^{n}(\eta+e^{-y_i}) \text{ for all } \eta \neq 0.
\end{aligned}$$

These are two polynomials in η of degree n, so they have the same roots. So

$$\{e^{-x_1}, e^{-x_2}, \cdots, e^{-x_n}\} = \{e^{-y_1}, e^{-y_2}, \cdots, e^{-y_n}\} \qquad (1.6.13)$$

and X and Y have the same order statistics. U is equivalent to $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$, which is minimal sufficient. □

2 Point Estimation

2.1 Estimation, Loss and Risk

Model: $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$.
Goal: To estimate $g(\theta) \in \mathbb{R}^k$ by an *estimator* $\delta(X) \in \mathbb{R}^k$ and measure error by *loss function* $L(\theta, d)$ such as $L(\theta, \delta) = \sum_{i=1}^{k} (g(\theta)_i - \delta_i)^2$.

Definition 2.1.1 (Unbiasedness)**.** $\delta(X)$ is *unbiased* for $g(\theta)$ if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ for *all* $\theta \in \Omega$.

Definition 2.1.2 (Risk)**.** The *risk* of $\delta(X)$ for estimating $g(\theta)$ under loss $L(\theta, \delta)$ is

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]. \tag{2.1.1}$$

Here, the expectation is with respect to $X \sim P_\theta$ and the risk is a function of θ.

Proposition 2.1.1. Let T be sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$. For any estimator $\delta(X)$ of $g(\theta)$, there exists a randomized estimator based only on $T(X)$ that has the same risk.

Proof. Generate X' from conditional distribution $X|T(X)$, which doesn't depend on θ. Estimate $g(\theta)$ by $\delta(X')$. □

Theorem 2.1.1 (Rao-Blackwell)**.** Let T be sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$. Given an estimator $\delta(X)$ for $g(\theta)$, define $\delta'(T(X))$ by

$$\delta'(t) = \mathbb{E}[\delta(X)|T(X) = t] \tag{2.1.2}$$

$\delta'(T(X))$ doesn't depend on θ becasue T is sufficient.

(i) If $L(\theta, \delta)$ is convex in d, then

$$R(\theta, \delta') \le R(\theta, \delta). \tag{2.1.3}$$

(ii) If $R(\theta, \delta) < \infty$, $L(\theta, \delta)$ is strictly convex in δ, and $\mathbb{P}_\theta[\delta'(T(X)) = \delta(X)] \neq 1$, then

$$R(\theta, \delta') < R(\theta, \delta). \tag{2.1.4}$$

Proof.
Jensen's Inequality: If f is convex, then

$$\mathbb{E}[f(X)] \ge f(\mathbb{E}[X]).$$

If f is strictly convex, X is not constant a.e., and $\mathbb{E}[f(X)] < \infty$, then

$$\mathbb{E}[f(X)] > f(\mathbb{E}[X]).$$

Apply Jensen's Inequality conditional on $T(x)$

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[L(\theta, \delta(X))] \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta[L(\theta, \delta(X))|T(X)]] \\ &\ge \mathbb{E}_\theta[L(\theta, \mathbb{E}_\theta[\delta(X)|T(X)])] \\ &= \mathbb{E}_\theta[L(\theta, \delta'(T(X)))] = R(\theta, \delta'). \end{aligned} \tag{2.1.5}$$

If $L(\theta, d)$ is strictly convex and $\mathbb{P}_\theta[\delta'(T(X)) = \delta(X)] \neq 1$, then we got strict inequality. □

Note 2.1.1. For convex loss, we should always estimate by function of sufficient statistic. $\delta'(T(X))$ is sometimes called the "Rao-Blackwellized estimate".

Example 2.1.1. Suppose $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$. To estimate $g(\theta) = \theta$, consider $\delta(X) = X_1$, $T(X) = X_1 + X_2 + \cdots + X_n$ is minimal sufficient. Consider

$$\begin{aligned}
&\delta'(t) = \mathbb{E}[X_1 | T(X) = X_1 + X_2 + \cdots + X_n = t] \\
&\text{By symmetry, this is the same as} \\
&\delta'(t) = \mathbb{E}[X_i | T(X) = X_1 + X_2 + \cdots + X_n = t] \text{ for every } i = 1, 2, \cdots, n \\
&n\delta'(t) = \sum_{i=1}^{n} \mathbb{E}[X_i | T(X) = X_1 + X_2 + \cdots + X_n = t] = t \\
&\delta'(t) = \frac{1}{n} t = \frac{1}{n} \sum_{i=1}^{n} X_i
\end{aligned} \tag{2.1.6}$$

Example 2.1.2. Suppose $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$. To estimate $g(\theta) = \mathbb{P}_\theta[X \le -2] = \Phi(-2 - \theta)$, our first guess $\delta(X) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le -2)$ is not a function of minimal sufficient statistics.

$$\begin{aligned}
\delta'(t) &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le -2) | \bar{X} = t \right] \\
&= \frac{1}{n} \mathbb{P}[X_i \le -2 | \bar{X} = t] \\
&= \frac{1}{n} \mathbb{P}[X_i - \bar{X} \le -2 - t | \bar{X} = t]
\end{aligned}$$

In this model, $X_i - \bar{X}$ and $\bar{X}$ are independent (Check $(X_i - \bar{X}, \bar{X})$ is bivariate normal, $Cov[X_i - \bar{X}, \bar{X}] = Cov[X_i, \bar{X}] - Var(\bar{X}) = \frac{1}{n} - \frac{1}{n} = 0$, and $X_i - \bar{X} \sim N(0, \frac{n-1}{n})$). So

$$\begin{aligned}
\delta'(t) &= \frac{1}{n} \mathbb{P}[X_i - \bar{X} \le -2 - t] \\
&= \frac{1}{n} \mathbb{P}\left[\sqrt{\frac{n}{n-1}} (X_i - \bar{X}) \le \sqrt{\frac{n}{n-1}} (-2 - t) \right] \\
&= \Phi\left(\sqrt{\frac{n}{n-1}} (-2 - t) \right) \\
&= \Phi\left(\sqrt{\frac{n}{n-1}} (-2 - \bar{X}) \right)
\end{aligned} \tag{2.1.7}$$

Note 2.1.2. If $\delta(X)$ is unbiased (i.e., $\mathbb{E}_\theta[\delta(X)] = g(\theta)$). Then

$$\mathbb{E}_\theta[\delta'(X)] = \mathbb{E}_\theta[\mathbb{E}[\delta(X) | T(X)]] = g(\theta). \tag{2.1.8}$$

So $\delta'(X)$ is also unbiased. Not true for "plug-in" estimate $\Phi(-2 - \bar{X})$.

Remark 2.1.1. In most exponential family models, as long as there is *some* unbiased estimate for $g(\theta)$, there is a *unique* unbiased estimator $\delta'(T(X))$ with minimum variance (or minimum risk for any convex loss), which may be obtained by Rao-Blackwellizing any unbiased estimator $\delta(X)$.

Definition 2.1.3 (Admissibility)**.** For the estimators $\delta_1,\ \delta_2$ if

(i) $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ at all $\theta \in \Omega$.

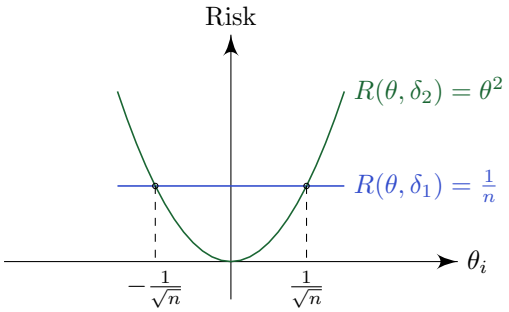
(ii) $R(\theta, \delta_1) < R(\theta, \delta_2)$ at some $\theta \in \Omega$.

Then, δ_1 dominates δ_2, and δ_2 is inadmissable. If no estimator dominates δ_2, then δ_2 is admissable.

Example 2.1.3. Suppose $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$. And there are two estimators of θ: $\delta_1(X) = \bar{X}$, $\delta_2(X) = 0$. Under loss $L(\theta, \delta) = (\theta - \delta)^2$,

$$\begin{aligned} R_1(\theta, \delta) &= \mathbb{E}_\theta[(\bar{X} - \theta)^2] = \frac{1}{n} \\ R_2(\theta, \delta) &= \mathbb{E}_\theta[(0 - \theta)^2] = \theta^2 \end{aligned} \tag{2.1.9}$$

δ_2 is "better" if $|\theta| \leq \frac{1}{\sqrt{n}}$.



Question 2.1.1. How to compare δ_1 and δ_2 when neither dominates the other?

Answer 2.1.1.

- Paradigm 1: Bayesian: Consider weight function $\pi : \Omega \to (0, \infty)$ where $\int_\Omega \pi(\theta) \mathrm{d}\theta = 1$ and compare average risk $\int_\Omega \pi(\theta) L(\theta, \delta) \mathrm{d}\theta$.
- Paradigm 2: Minimax: Compare by the worst-case risk $\sup_{\theta \in \Omega} R(\theta, \delta)$.

2.2 Bayesian Estimation

Recall: For estimating $g(\theta)$ under a loss $L(\theta, \delta)$, the *risk* of an estimator $\delta(X)$ is $R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$.

Definition 2.2.1 (Bayes estimator)**.** For a model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, a *prior distribution* Λ is a probability distribution over the parameter space Ω. The *Bayes risk* of an estimator $\delta(X)$ with respect to Λ is

$$r(\Lambda, \delta) = \int_\Omega R(\theta, \delta) \mathrm{d}\Lambda(\theta). \tag{2.2.1}$$

An estimator $\delta(X)$ such that

$$r(\Lambda, \delta) \leq r(\Lambda, \delta') \quad \text{for all other estimators } \delta' \tag{2.2.2}$$

is a *Bayes estimator*.

Note 2.2.1. If Λ has density $\pi(\theta)$ on $\Omega \subseteq \mathbb{R}^k$, then

$$r(\Lambda, \delta) == \int_\Omega \pi(\theta) R(\theta, \delta) \mathrm{d}\theta. \tag{2.2.3}$$

The alternative interpretation is that Λ encodes our belief, state of mind, information from previous experiments, $\cdots$

Remark 2.2.1. Let $\Theta \sim \Lambda$ be random, $X|\Theta = \theta \sim P_\theta$. Then

$$r(\Lambda, \delta) = \mathbb{E}[L(\Theta, \delta(X))], \tag{2.2.4}$$

where $\mathbb{E}$ is over both Θ and X. And we can write it as

$$\begin{aligned} r(\Lambda, \delta) &= \mathbb{E}[\mathbb{E}[L(\Theta, \delta(X))|\Theta]] \\ &= \int_\Omega \mathbb{E}_\theta[L(\Theta, \delta(X))] \mathrm{d}\Lambda(\theta) \\ &= \int_\Omega R(\theta, \delta(X)) \mathrm{d}\Lambda(\theta). \end{aligned} \tag{2.2.5}$$

or alternatively as

$$\begin{aligned} r(\Lambda, \delta) &= \mathbb{E}[\mathbb{E}[L(\Theta, \delta(X))|X]] \\ &= \int_{\mathcal{X}} \mathbb{E}[L(\Theta, \delta(X))|X = x] \mathrm{d}P^X(x). \end{aligned} \tag{2.2.6}$$

Here, P^X is the marginal distribution of X. If (Θ, X) have joint density $\pi(\theta, x)$, then P^X has density

$$\pi(x) = \int_\Omega \pi(\theta, x) \mathrm{d}\theta. \tag{2.2.7}$$

Definition 2.2.2 (Posterior distribution)**.** The distribution of Θ given $X = x$ is the *posterior distribution* of Θ.

Notation 2.2.1.

$\pi(\theta)$ is the prior distribution
$\pi(x|\theta) \equiv p_\theta(x)$
$\pi(\theta|x)$ is the posterior distribution
$\pi(x)$ is the marginal distribution of X.

Theorem 2.2.1. Suppose $\Theta \sim \Lambda$, $X \sim P_\theta$ conditional on $\Theta = \theta$, $L(\theta, d)$ is non-negative and for all $x \in \mathcal{X}$ or for all $x \notin N$ where $P^X(N) = 0$, there exists a value $\delta_\Lambda(x)$ which minimizes $\mathbb{E}[L(\Theta, \delta(X))|X = x]$. Then

(i) δ_Λ is a Bayes estimator for Λ.

(ii) If $L(\theta, d)$ is strictly convex in d, then δ_Λ is the unique Bayes estimator $p^X - a.e.$ In other words, if δ'_Λ is another Bayes estimator, then $\delta_\Lambda(x) = \delta'_\Lambda(x)$ for all $x \notin N$, where $P^X(N) = 0$.

Proof.
For minimization: For any other $\delta(X)$,

$$\mathbb{E}[L(\Theta, \delta_\Lambda(X))|X = x] \leq \mathbb{E}[L(\Theta, \delta(X))|X = x] \quad \text{for } P^X - a.e.\ x \in \mathcal{X}. \tag{2.2.8}$$

Take expectations with respect to $X \sim P^X$, we have

$$r(\Lambda, \delta_\Lambda) \leq r(\Lambda, \delta). \tag{2.2.9}$$

For uniqueness: Suppose $\delta'_\Lambda(x) \neq \delta_\Lambda(x)$ for any $x \in S$, where $P^X(S) > 0$. L is strictly convex implies that if $\delta_\Lambda(x)$ minimizes $\mathbb{E}[L(\Theta, \delta(X))|X = x]$, then it is the unique statistic for minimization. Therefore,

$$\mathbb{E}[L(\Theta, \delta_\Lambda(X))|X = x] < \mathbb{E}[L(\Theta, \delta'_\Lambda(X))|X = x] \quad \text{for } \ x \in S. \tag{2.2.10}$$

Take expectations with respect to $X \sim P^X$, we have

$$r(\Lambda, \delta_\Lambda(X)) < r(\Lambda, \delta'_\Lambda(X)). \tag{2.2.11}$$

So $\delta'_\Lambda(x)$ is not Bayes. □

Corollary 2.2.1.

(i) If $L(\theta, \delta) = (g(\theta) - \delta)^2$, then the unique Bayes is

$$\delta_\Lambda(x) = \mathbb{E}[g(\Theta)|X = x]. \tag{2.2.12}$$

(ii) If $L(\theta, \delta) = w(\theta)(g(\theta) - \delta)^2$ for $w : \Omega \to (0, \infty)$, then the unique Bayes is

$$\delta_\Lambda(x) = \frac{\mathbb{E}[w(\Theta)g(\Theta)|X = x]}{\mathbb{E}[w(\Theta)|X = x]}. \tag{2.2.13}$$

(iii) If $L(\theta, \delta) = |\theta - \delta|$, then the Bayes is

$$\delta_\Lambda(x) = \text{ any median of } \Theta|X = x. \tag{2.2.14}$$

(iv) If $L(\theta, \delta) = \begin{cases} 0 & |\delta - \theta| \leq c \\ 1 & |\delta - \theta| > c \end{cases}$ for fixed $c > 0$, then the Bayes is

$$\begin{aligned} \delta_\Lambda(x) =& \text{midpoint of any interval } I \text{ of length } 2c \text{ which maximizes} \\ & \mathbb{P}[\Theta \in I|X = x] \end{aligned} \tag{2.2.15}$$

Proof. If $Y \sim P$, then

(i) δ minimizing $\mathbb{E}[(Y - \delta)^2]$ is

$$\delta = \mathbb{E}[Y]. \tag{2.2.16}$$

(ii) δ minimizing $\mathbb{E}[(Y - \delta)^2 \cdot w(Y)]$ is

$$\delta = \frac{\mathbb{E}[w(Y) \cdot Y]}{\mathbb{E}[w(Y)]}. \tag{2.2.17}$$

Since this is Minimizing quadratic over δ.

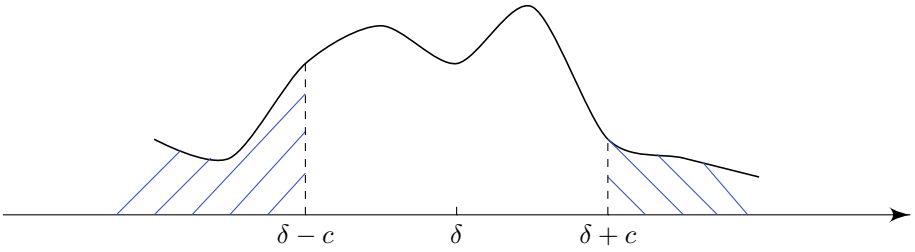
$$\mathbb{E}[(Y - \delta)^2 \cdot w(Y)] = \mathbb{E}[Y^2 \cdot w(Y)] - 2\delta\mathbb{E}[Y \cdot w(Y)] + \delta^2\mathbb{E}[w(Y)].$$

(iii) δ minimizing $\mathbb{E}[|Y-\delta|]$ is

$$\delta = \text{ any median of } P. \tag{2.2.18}$$

(iv) δ minimizing $\mathbb{P}[|Y-\delta| > c] \iff$ maxmizing $\mathbb{P}[|Y-\delta| \le c] = \mathbb{P}[Y \in [\delta-c, \delta+c]]$ is

$$\delta = \text{midpoint of such an interval } I. \tag{2.2.19}$$



Apply these conclusions to P = poisterior distribution of $\Theta|X=x$. □

Example 2.2.1. Let $X_1, X_2, \cdots, X_n, \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. Suppose the prior distribution is $\Theta \sim \Lambda = Beta(a,b)$. So

$$\pi(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}, \quad \pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

Then

$$\pi(\theta, x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{\sum_{i=1}^{n} x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1}, \text{ and}$$

$$\pi(\theta|x) = \frac{\pi(\theta,x)}{\pi(x)} = C(x,a,b,n)\theta^{\sum_{i=1}^{n} x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1}.$$

$C(x,a,b,n)$ must be normalizing constant such that

$$\int_0^1 \pi(\theta|x)\mathrm{d}\theta = 1.$$

Therefore,

$$C(x,a,b,n) = \frac{\Gamma(n+a+b)}{\Gamma(\sum_{i=1}^{n} x_i + a)\Gamma(n - \sum_{i=1}^{n} x_i + b)},$$

$$\Theta|X = x \sim Beta\left(\sum_{i=1}^{n} x_i + a, n - \sum_{i=1}^{n} x_i + b\right) \equiv Beta(a', b').$$

$Beta(a,b)$ is a *conjugate prior* for Θ.

Consider estimating θ under $L(\theta, \delta) = (\theta - \delta)^2$.

$$\begin{aligned}\delta_\Lambda(x) &= \mathbb{E}[\Theta|X = x] = \int_0^1 \theta \cdot \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}\theta^{a'-1}(1-\theta)^{b'-1}\mathrm{d}\theta = \frac{a'}{a'+b'},\\ \delta_\Lambda(x) &= \frac{\sum_{i=1}^n x_i + a}{n+a+b} = \left(\frac{a+b}{n+a+b}\right)\cdot\frac{a}{a+b} + \left(\frac{n}{n+a+b}\right)\cdot\frac{\sum_{i=1}^n x_i}{n}.\end{aligned} \tag{2.2.20}$$

Interpretation: a = prior number of $1's$, b = prior number of $0's$, $a+b$ = prior sample size, $\frac{a}{a+b}$ = prior mean; $\sum_{i=1}^n x_i$ = number of $1's$ in sample, $n - \sum_{i=1}^n x_i$ = number of $0's$ in sample, n = sample size, $\frac{\sum_{i=1}^n x_i}{n}$ = sample mean.

Note: Recover $\bar{X}$ under the *improper prior* $a = 0,\ b = 0,\ \pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$. This is not a probability distribution as

$$\int_0^1 \theta^{-1}(1-\theta)^{-1}\mathrm{d}\theta \to \infty.$$

However, the posterior is proper as long as $\sum_{i=1}^n x_i \notin \{0, n\}$. $\bar{X}$ is called *generalized Bayes*. Alternatively, $\bar{X} = \lim_{a,b\to 0}\delta_\Lambda(X)$. $\bar{X}$ is a *limit of* (proper) Bayes estimators.

Consider estimating θ under $L(\theta, \delta) = \frac{(\theta-\delta)^2}{\theta(1-\theta)}$.

$$\delta_\Lambda(X) = \frac{\mathbb{E}[\frac{1}{\Theta(1-\Theta)}\cdot\Theta|X=x]}{\mathbb{E}[\frac{1}{\Theta(1-\Theta)}|X=x]} = \frac{\frac{a'+b'-1}{b'-1}}{\frac{(a'+b'-1)(a'+b'-2)}{(a'-1)(b'-1)}} = \frac{a'-1}{a'+b'-2} = \frac{\sum_{i=1}^n x_i + a - 1}{n+a+b-2}. \tag{2.2.21}$$

Example 2.2.2. Let $X_1, X_2, \cdots, X_n, \overset{\text{i.i.d.}}{\sim} Poisson(\theta)$. Suppose the prior distribution is $\Theta \sim \Lambda = Gamma(a, b)$. So

$$\pi(x|\theta) = \prod_{i=1}^n \frac{\theta^{x_i}e^{-\theta}}{x_i!}, \quad \pi(\theta) = \frac{1}{\Gamma(a)b^a}\theta^{a-1}e^{-\frac{\theta}{b}}.$$

Then

$$\pi(\theta|x) \propto \pi(\theta, x) \propto \theta^{\sum_{i=1}^n x_i + a - 1}e^{-n\theta - \frac{\theta}{b}}.$$

So

$$\Theta|X = x \sim Gamma\left(\sum_{i=1}^n x_i + a, \frac{1}{n+\frac{1}{b}}\right).$$

To estimate θ under $L(\theta, d) = (\theta - d)^2$:

$$\delta_\Lambda(x) = \mathbb{E}[\Theta|X=x] = \frac{\sum_{i=1}^n x_i + a}{n+\frac{1}{b}} = \left(\frac{\frac{1}{b}}{n+\frac{1}{b}}\right)\cdot ab + \left(\frac{n}{n+\frac{1}{b}}\right)\cdot\frac{\sum_{i=1}^n x_i}{n} \tag{2.2.22}$$

Interpretation: ab = prior sample mean; $\frac{\sum_{i=1}^n x_i}{n}$ = sample mean.

Example 2.2.3. Let $X_1, X_2, \cdots, X_n, \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$, σ^2 known constant. Suppose the prior distribution is $\Theta \sim N(\mu, \tau^2)$.So

$$\pi(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right), \quad \pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right).$$

Then

$$\begin{aligned}
\pi(\theta|x) &\propto \pi(\theta, x) \\
&\propto \exp\left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} \cdot \theta - \frac{n\theta^2}{2\sigma^2} + \frac{\mu}{\tau^2} \cdot \theta - \frac{\theta^2}{2\tau^2}\right). \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\left(\theta - \frac{\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)^2\right).
\end{aligned}$$

So

$$\Theta|X = x \sim N\left(\frac{\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right).$$

To estimate θ under $L(\theta, d) = (\theta - d)^2$:

$$\delta_\Lambda(x) = \mathbb{E}[\Theta|X = x] = \frac{\frac{\sigma^2}{\tau^2}}{n + \frac{\sigma^2}{\tau^2}} \cdot \mu + \frac{n}{n + \frac{\sigma^2}{\tau^2}} \cdot \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.2.23}$$

Interpretation: $\frac{\sigma^2}{\tau^2}$ = prior sample size, μ = prior sample mean; n = sample size, $\frac{\sum_{i=1}^{n} x_i}{n}$ = sample mean. $\frac{\sum_{i=1}^{n} x_i}{n} = \bar{X} = \lim_{\tau \to \infty} \delta_\Lambda(X)$ (for my fixed μ.)

Theorem 2.2.2. Consider estimating $g(\theta)$ under squared error loss: $L(\theta, d) = (\theta - d)^2$. No unbiased estimate $\delta(X)$ can be a Bayes estimate (for a proper prior Λ) unless $g(\theta)$ is perfectly estimable under Λ

$$\mathbb{P}[\delta(X) = g(\Theta)] = 1, \tag{2.2.24}$$

Proof. Suppose $\delta(X)$ is Bayes for Λ: $\delta(X) = \mathbb{E}[g(\Theta)|X]$. Then

$$\mathbb{E}[g(\Theta)\delta(X)] = \mathbb{E}[\mathbb{E}[g(\Theta)\delta(X)|X]] = \mathbb{E}[\delta(X)\mathbb{E}[g(\Theta)|X]] = \mathbb{E}[\delta(X)^2].$$

Now suppose δ is unbiased: $\mathbb{E}[\delta(X)|\Theta] = g(\Theta)$. Then

$$\mathbb{E}[g(\Theta)\delta(X)] = \mathbb{E}[\mathbb{E}[g(\Theta)\delta(X)|\Theta]] = \mathbb{E}[g(\Theta)\mathbb{E}[\delta(X)|\Theta]] = \mathbb{E}[g(\Theta)^2].$$

So

$$\begin{aligned}
&\mathbb{E}[(\delta(X) - g(\Theta))^2] = \mathbb{E}[\delta(X)^2] + \mathbb{E}[g(\Theta)^2] - 2\mathbb{E}[g(\Theta)\delta(X)]. \\
&\mathbb{P}[\delta(X) = g(\Theta)] = 1.
\end{aligned} \tag{2.2.25}$$

□

Example 2.2.4. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \pi(x|\eta)$, where

$$\pi(x|\eta) = \exp\left(\sum_{j=1}^{s} \eta_j T_j(x) - A(\eta)\right) h(x).$$

Take a conjugate prior

$$\pi(\eta) = c(k, \mu_1, \mu_2, \cdots, \mu_s) \exp\left(\sum_{j=1}^{s} k\mu_j \eta_j - kA(\eta)\right).$$

Then

$$\pi(\eta|x) \propto \exp\left(\sum_{j=1}^{s} \eta_j \left(k\mu_j + \sum_{i=1}^{n} T_j(x_i)\right) - (k+n)A(\eta)\right).$$

Interpretation: k = prior sample size, μ_j = prior mean of T_j.

(i) Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. Then

$$\pi(x_i|\theta) = \exp\left(x_i \log\frac{\theta}{1-\theta} + \log(1-\theta)\right)$$
$$\eta = \log\left(\frac{\theta}{1-\theta}\right),\ A(\eta) = \log(1+e^{\eta}).$$

Take a conjugate prior

$$\pi(\eta) \propto \exp(k\mu\eta - k\log(1+e^{\eta})).$$

Change variables:

$$\begin{aligned}\tilde{\pi}(\theta) &\propto \pi(\eta(\theta)) \cdot \frac{\mathrm{d}\eta}{\mathrm{d}\theta} \\ &= \left(\frac{\theta}{1-\theta}\right)^{k\mu} \left(\frac{1}{1-\theta}\right)^{-k} \cdot \frac{1}{\theta(1-\theta)} \\ &= \theta^{k\mu-1}(1-\theta)^{k(1-\mu)-1} \sim Beta(k\mu, k(1-\mu)).\end{aligned} \tag{2.2.26}$$

(ii) Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Poission(\theta)$. Then

$$\pi(x_i|\theta) = \exp(x_i \log\theta - \theta) \cdot \frac{1}{x_i!}$$
$$\eta = \log\theta,\ A(\eta) = e^{\eta}.$$

Then

$$\pi(\eta) \propto \exp(k\mu\eta - ke^{\eta}).$$

Change variables:

$$\begin{aligned}\tilde{\pi}(\theta) &\propto \pi(\eta(\theta)) \cdot \frac{\mathrm{d}\eta}{\mathrm{d}\theta} \\ &= \theta^{k\mu} e^{-k\theta} \cdot \frac{1}{\theta} \\ &= \theta^{k\mu-1} e^{-k\theta} \sim Gamma\left(k\mu, \frac{1}{k}\right).\end{aligned} \tag{2.2.27}$$

2.3 Empirical and Hierarchical Bayes

Question 2.3.1. How to set prior in the absence of strong prior information?

Answer 2.3.1.

(i) Uninformative prior: convey little information

Example 2.3.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. We saw prior distribution $Beta(0,0)$

$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$$

recovers $\bar{X}$ as Bayes estimator for squared-error risk.

Bayes prior: Uniform prior distribution $\pi(\theta) = 1$, which means all θ are equally probable. Same as $Beta(1,1)$.
Note: If we change parameters to $\eta = \log(\frac{\theta}{1-\theta})$, we got

$$\tilde{\pi}(\eta) = \pi(\theta(\eta)) \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\eta} = \left(\frac{1}{1+e^{\eta}}\right)^2. \tag{2.3.1}$$

So not all η are equally probable.
Jeffreys prior: Let $I(\theta)$ be the *Fisher information*.

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(x)\right)^2\right]$$

Take prior distribution

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

Let $p(x|\eta)$ be the density parametrized by η. Then

$$\begin{aligned}\tilde{\pi}(\eta) = \pi(\theta(\eta)) \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\eta} &\propto \sqrt{I(\theta)\cdot\left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2} \\ &= \sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(x)\cdot\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2\right]} \\ &= \sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\eta}\log p(x|\eta)\right)^2\right]} = \sqrt{I(\eta)}.\end{aligned} \tag{2.3.2}$$

This is the prior we would have used, had we started in the parametrization η and taken the Jeffreys prior which is is parametrization invariant:

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

For $Bernoulli(\theta)$

$$\begin{aligned}&\frac{\partial}{\partial\theta}\log p_\theta(x) = \frac{\partial}{\partial\theta}[x\log\theta + (1-x)\log(1-\theta)] = \frac{x}{\theta} - \frac{1-x}{1-\theta}. \\ &I(\theta) = \mathbb{E}_\theta\left[\left(\frac{x}{\theta} - \frac{1-x}{1-\theta}\right)^2\right] = \frac{1}{\theta(1-\theta)}. \\ &\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}.\end{aligned} \tag{2.3.3}$$

$Beta\left(\frac{1}{2}, \frac{1}{2}\right)$ is the Jeffreys prior.

Example 2.3.2. Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$.

$$\begin{aligned} &\frac{\partial}{\partial\theta}\log p_\theta(x) = \frac{\partial}{\partial\theta}\left[-\frac{(x-\theta)^2}{2} - \log\sqrt{2\pi}\right] = x-\theta. \\ &I(\theta) = \mathbb{E}_\theta\left[(x-\theta)^2\right]. \\ &\pi(\theta) \propto 1. \end{aligned} \tag{2.3.4}$$

$\pi(\theta) = 1$ is the Jeffreys prior, and the improper uninformative prior on $\mathbb{R}$.

(ii) Empirical Bayes: estimate prior from data

Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \pi(x|\theta)$, $\Theta \sim \pi(\theta|\gamma)$. Here, γ parametrizes the prior. The marginal density of $X = (X_1, X_2, \cdots, X_n)$ is

$$\pi(x|\gamma) = \int \prod_{i=1}^n \pi(x_i|\theta)\cdot\pi(\theta|\gamma)\mathrm{d}\theta.$$

Based on $\pi(x|\gamma)$, we estimate γ by $\hat{\gamma}(x)$, then we use this for Bayes estimation of θ.

Example 2.3.3. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, $\Theta \sim N(0, \tau^2)$. Under squared-error loss, the Bayes estimate is

$$\mathbb{E}[\Theta|X] = \frac{n}{n+\frac{1}{\tau^2}}\bar{X}.$$

To estimate τ^2

$$\begin{aligned} &\pi(x_1, x_2, \cdots, x_n|\tau^2) \\ &= \int \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} \cdot \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\theta^2}{2\tau^2}}\,\mathrm{d}\theta \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{2\pi\tau^2}} \int \exp\left(-\frac{1}{2}\left(n+\frac{1}{\tau^2}\right)\theta^2 + \sum_{i=1}^n x_i\cdot\theta - \frac{1}{2}\sum_{i=1}^n x_i^2\right)\mathrm{d}\theta \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{2\pi\tau^2}} \int \exp\left(-\frac{1}{2}\left(n+\frac{1}{\tau^2}\right)\left(\theta - \frac{\sum_{i=1}^n x_i}{n+\frac{1}{\tau^2}}\right)\right)^2 \mathrm{d}\theta \\ &\cdot \exp\left(\frac{1}{2}\left(n+\frac{1}{\tau^2}\right)\left(\frac{\sum_{i=1}^n x_i}{n+\frac{1}{\tau^2}}\right)^2 - \frac{1}{2}\sum_{i=1}^n x_i^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{1+n\tau^2}} \cdot \exp\left(\frac{1}{2}\frac{n^2}{n+\frac{1}{\tau^2}}\cdot\bar{x}^2 - \frac{1}{2}\sum_{i=1}^n x_i^2\right) \end{aligned}$$

Here, $\bar{x}^2$ is a sufficient statistic.

Note: $\bar{X}|\theta \sim N(\theta, \frac{1}{n})$, $\Theta \sim N(0, \tau^2)$. So $\bar{X} = \Theta + W$ where $W \sim N(0, \frac{1}{n})$ is independent of Θ.

$$\bar{X} \sim N\left(0, \tau^2 + \frac{1}{n}\right).$$

$$\mathbb{E}[\bar{X}^2] = \tau^2 + \frac{1}{n},$$

So $\hat{\tau}^2 = \bar{X}^2 - \frac{1}{n}$ is unbiased for τ^2. And we can check that this is also the maximum likelihood estimate.
Therefore, the empirical Bayes estimate of θ is

$$\delta(X) = \frac{n}{n + \frac{1}{\hat{\tau}^2}} \bar{X} = \left(1 - \frac{1}{n\bar{X}^2}\right) \bar{X}. \tag{2.3.5}$$

(iii) Hierarchical Bayes: put a hyper prior on γ
Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \pi(x|\theta),\ \Theta \sim \pi(\theta|\gamma),\ \gamma \sim \pi(\gamma)$.
Note: This is equivalent to putting the marginal prior

$$\pi(\theta) = \int \pi(\theta|\gamma)\pi(\gamma)\mathrm{d}\gamma$$

on Θ. But, it's often easier to define a robust, heavy-tailed prior this way.

Example 2.3.4. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, $\Theta \sim N(0, \tau^2)$, $\frac{1}{\tau^2} \sim Gamma(\frac{\nu}{2}, \frac{2}{\nu})$. It implies marginal prior $\pi(\theta)$ is student-t distribution with ν degrees of freedom.
Note: $\pi(\theta|\tau^2)$ is conjugate for $\pi(x|\theta)$, and $\pi(\tau^2)$ is conjugate for $\pi(\theta(\tau^2))$. However, $\pi(\theta)$ is not conjugate for $\pi(x|\theta)$. There is no closed-form expression for $\mathbb{E}[\Theta|X]$. But there are closed expression for the conditional laws of Θ, τ^2. Let $\omega = \frac{1}{\tau^2}$

$$\begin{aligned} \Theta|X, \omega &\sim N\left(\frac{n\bar{X}}{n+\omega}, \frac{1}{n+\omega}\right) \\ \omega|\Theta, X &\sim Gamma\left(\frac{\nu+1}{2}, \frac{2}{\nu+\theta^2}\right). \end{aligned} \tag{2.3.6}$$

Gibbs-sampling approach to estimate θ:

- Initialize $\omega^{(0)}$.
- For $t = 1, 2, 3, \cdots$
  - Sample $\theta^{(t)}$ from conditional distribution $\Theta|X, \omega^{(t-1)}$.
  - Sample $\omega^{(t)}$ from conditional distribution $\omega|\Theta^{(t)}, X$.

The output is $(\theta^{(1)}, \omega^{(1)}), (\theta^{(2)}, \omega^{(2)}), \cdots$
This algorithm guarantees the output pairs form a Markov chain with stationary distribution $\Theta, \omega|X$.
To compute $\mathbb{E}[\Theta|X]$, we can use the following two estimators:

- $\delta_1 = \frac{1}{T} \sum_{t=1}^{T} \theta^{(t)}$.
- $\delta_2 = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\Theta|X, \omega^{(t-1)}] = \frac{1}{T} \sum_{t=1}^{T} \frac{n\bar{X}}{n+\omega^{(t-1)}}$.

Note: δ_2 is Rao-Blackwellized and often has lower variance than δ_1.

2.4 Minimax Estimation

Setup: To estimate $g(\theta)$ in model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ under loss $L(\theta, d)$.
Goal: To find δ minimizing the worst-case risk

$$\begin{aligned} \bar{r}(\delta) &= \sup_{\theta \in \Omega} R(\theta, \delta), \\ R(\theta, \delta) &= \mathbb{E}_\theta[L(\theta, \delta(X))]. \end{aligned} \tag{2.4.1}$$

Definition 2.4.1 (Minimax)**.** $\delta(X)$ is a *minimax* estimator if

$$\bar{r}(\delta) = \inf_{\delta'} \bar{r}(\delta') = \inf_{\delta'} \sup_{\theta \in \Omega} R(\theta, \delta'), \tag{2.4.2}$$

where $\inf_{\delta'}$ is taken over *all* possible estimators. The value of the risk

$$R(\Omega) = \inf_{\delta} \sup_{\theta \in \Omega} R(\theta, \delta) \tag{2.4.3}$$

is the *minimax risk* of this estimation problem.

Remark 2.4.1. For any estimator δ and prior Λ,

$$\begin{aligned} \bar{r}(\delta) &\geq \int R(\theta, \delta) \mathrm{d}\Lambda(\theta) \equiv r(\Lambda, \delta). \\ &\text{worst-case risk} \geq \text{average risk.} \end{aligned} \tag{2.4.4}$$

Take minimum over δ

$$R(\Omega) = \inf_{\delta} \bar{r}(\delta) \geq \inf_{\delta} r(\Lambda, \delta) = r(\Lambda, \delta_\Lambda) \equiv B(\Lambda), \tag{2.4.5}$$

where δ_Λ is a Bayes estimator for Λ and $B(\Lambda)$ is its Bayes risk. This holds for all priors Λ on Ω, so

$$R(\Omega) \geq \sup_{\Lambda} B(\Lambda). \tag{2.4.6}$$

This is a *minimax lower bond*.

Idea: δ is minimax if this lower bond is tight.

Definition 2.4.2 (Least-favorable prior)**.** Λ is a *least-favorable prior* if

$$B(\Lambda) = \sup_{\Lambda'} B(\Lambda'). \tag{2.4.7}$$

$\Lambda_1, \Lambda_2, \cdots$ is a *least-favorable sequence* of priors if

$$\lim_{i \to \infty} B(\Lambda_i) = \sup_{\Lambda'} B(\Lambda'). \tag{2.4.8}$$

Theorem 2.4.1.

(i) (Proper Bayes estimators): Let δ be an estimator, and Λ is a prior such that

$$\bar{r}(\delta) = B(\Lambda) \tag{2.4.9}$$

Then

- δ is minimax.
- δ is a Bayes estimator for Λ

$$\bar{r}(\delta) = r(\Lambda, \delta) = \int R(\theta, \delta) \mathrm{d}\Lambda(\theta).$$

- Λ is a least-favorable prior.

(ii) (More general estimators): Let δ be an estimator, and $\Lambda_1, \Lambda_2, \cdots$ is a sequence of priors such that

$$\bar{r}(\delta) = \lim_{i \to \infty} B(\Lambda_i). \tag{2.4.10}$$

Then

- δ is minimax.
- $\Lambda_1, \Lambda_2, \cdots$ is a least-favorable sequence of priors.

Proof.

(i) (Proper Bayes estimators):

- Let δ' be any other estimator. By (2.4.4)

$$\bar{r}(\delta') \geq r(\Lambda, \delta') \geq B(\Lambda) = \bar{r}(\delta). \tag{2.4.11}$$

So δ is minimax. Applying this with $\delta' = \delta$, we must have

$$\begin{aligned} &\bar{r}(\delta) = r(\Lambda, \delta), \\ &r(\Lambda, \delta) = B(\Lambda) \quad \text{(i.e., } \delta \text{ is Bayes for } \Lambda). \end{aligned} \tag{2.4.12}$$

- Consider any prior Λ'. By (2.4.4)

$$B(\Lambda') \leq r(\Lambda', \delta) \leq \bar{r}(\delta) = B(\Lambda). \tag{2.4.13}$$

So Λ is a least-favorable prior.

(ii) (More general estimators):

- Let δ' be any other estimator. By (2.4.4)

$$\bar{r}(\delta') \geq r(\Lambda_i, \delta') \geq B(\Lambda_i) \quad \text{for all } i. \tag{2.4.14}$$

So

$$\bar{r}(\delta') \geq \lim_{i \to \infty} B(\Lambda_i) = \bar{r}(\delta). \tag{2.4.15}$$

So δ is minimax.

– Consider any prior Λ'. By (2.4.4)

$$B(\Lambda') \le r(\Lambda', \delta) \le \bar{r}(\delta) = \lim_{i\to\infty} B(\Lambda_i). \tag{2.4.16}$$

So $\Lambda_1, \Lambda_2, \cdots$ is a least-favorable sequence of priors.

□

Remark 2.4.2.

(i) In Theorem 2.4.1 (i), if δ_Λ is the *unique* Bayes estimate for Λ, then it is unique minimax:

$$\bar{r}(\delta') \ge r(\Lambda, \delta') > r(\Lambda, \delta_\Lambda) = \bar{r}(\delta_\Lambda). \quad \text{for any other } \delta'. \tag{2.4.17}$$

We cannot get a similar statement from Theorem 2.4.1 (ii), and in general there may be *many* minimax estimators even if each Bayes estimate δ_{Λ_i} is unique.

(ii) In Theorem 2.4.1 (i), if δ_Λ has constant risk is constant over a set of θ with probability 1 under Λ, then it is minimax as (2.4.4) holds with equality.

Example 2.4.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$.

(i) We want to find a minimax estimator for θ under $L(\theta, \delta) = (\theta - \delta)^2$.

– First guess: $\delta(X) = \bar{X}$. Is this minimax?

$$R(\theta, \delta) = \mathbb{E}_\theta[(\bar{X} - \theta)^2] = Var_\theta[\bar{X}] = \frac{\theta(1-\theta)}{n}.$$

This doesn't see minimax because the risk is not constant and the worst case is achieved at single point $\theta = \frac{1}{2}$. Can we find an estimator with constant risk?

– Consider conjugate Beta prior: $\Theta \sim Beta(a, b)$. By Example 2.2.1 and (2.2.20), the posterior distribution for Θ is $Beta(a+n\bar{X}, b+n(1-\bar{X}))$, the Bayes estimator is

$$\delta_\Lambda = \mathbb{E}[\Theta|X] = \frac{a + n\bar{X}}{a + b + n}. \tag{2.4.18}$$

The risk function $R(\theta, \delta_\Lambda)$ for this Bayes estimator is

$$\begin{aligned} R(\theta, \delta_\Lambda) &= \mathbb{E}_\theta[(\delta_\Lambda - \theta)^2] \\ &= \mathbb{E}_\theta[((\delta_\Lambda - \mathbb{E}_\theta[\delta_\Lambda]) + (\mathbb{E}_\theta[\delta_\Lambda] - \theta))^2] \\ &= Var(\delta_\Lambda) + \mathbb{E}_\theta[(\mathbb{E}_\theta[\delta_\Lambda] - \theta)^2] \\ &= \frac{n\theta(1-\theta)}{(a+b+n)^2} + \frac{(a - \theta(a+b))^2}{(a+b+n)^2} \\ &= \frac{1}{(a+b+n)^2}[n\theta(1-\theta) + (a - \theta(a+b))^2] \\ &= \frac{1}{(a+b+n)^2}[((a+b)^2 - n)\theta^2 + (n - 2a(a+b))\theta + a^2] \end{aligned} \tag{2.4.19}$$

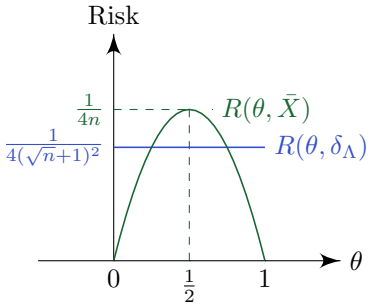
To get constant risk, we want:

$$\mu(A) = \left\{ \begin{array}{l} (a+b)^2 - n = 0 \\ n - 2a(a+b) = 0 \end{array} \right\} \Longrightarrow a = b = \frac{\sqrt{n}}{2}$$

We've shown if $\Lambda = Beta(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2})$, then δ_Λ has constant risk

$$r(\Lambda, \delta_\Lambda) = \frac{a^2}{(a+b+n)^2} = \frac{1}{4(1+\sqrt{n})^2} \tag{2.4.20}$$

By Theorem 2.4.1 (i), δ_Λ is minimax. In fact, since δ_Λ is the unique Bayes estimator for Λ, it is also unique minimax.



Note: Risk is better only in a small interval around $\theta = \frac{1}{2}$. For large n, we would still prefer the maximum likelihood estimation.

(ii) We want to find a minimax estimator for θ under $L(\theta, \delta) = \frac{(\theta-\delta)^2}{\theta(1-\theta)}$. Then

$$R(\theta, \bar{X}) = \mathbb{E}_\theta \left[\frac{(\theta - \bar{X})^2}{(1-\theta)\theta} \right] = \frac{1}{n}. \tag{2.4.21}$$

This is constant now in θ. By Example 2.2.1 and (2.2.21), the Bayes estimator is

$$\delta_\Lambda = \frac{a + n\bar{X} - 1}{a + b + n - 2}. \tag{2.4.22}$$

For $a = b = 1$, $\bar{X}$ is the Bayes estimator. so

$$\bar{r}(\bar{X}) = r(\Lambda, \bar{X}) = B(\Lambda) = \frac{1}{n}. \tag{2.4.23}$$

So $\bar{X}$ is now (unique) minimax estimator.

Example 2.4.2. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$, σ^2 is known. We guess that $\bar{X}$ is minimax, and the risk is

$$R(\theta, \bar{X}) = \mathbb{E}_\theta[(\bar{X} - \theta)^2] = \frac{\sigma^2}{n}. \tag{2.4.24}$$

This is constant in θ. But $\bar{X}$ cannot be Bayes for any proper prior Ω. So we want to construct a sequence of priors where $B(\Lambda_i) \to \frac{\sigma^2}{n}$.

Consider conjugate prior $\Lambda = N(0, \tau^2)$. By Example 2.2.3 and (2.2.23), the posterior distribution for Θ is $N\left(\frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\bar{X}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)$, the Bayes estimator is

$$\delta_\Lambda(x) = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\bar{X}. \tag{2.4.25}$$

So the Bayes risk is

$$\begin{aligned} B(\Lambda) &= \mathbb{E}[(\delta_\Lambda(x) - \Theta)^2] \\ &= \int \mathbb{E}[(\delta_\Lambda(x) - \Theta)^2 | X = x] \mathrm{d}P^X(x) \\ &= \int Var[\Theta | X = x] \mathrm{d}P^X(x) \\ &= \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} \\ \text{Let } \tau^2 \to \infty, \quad B(\Lambda) &\to \frac{\sigma^2}{n}. \end{aligned} \tag{2.4.26}$$

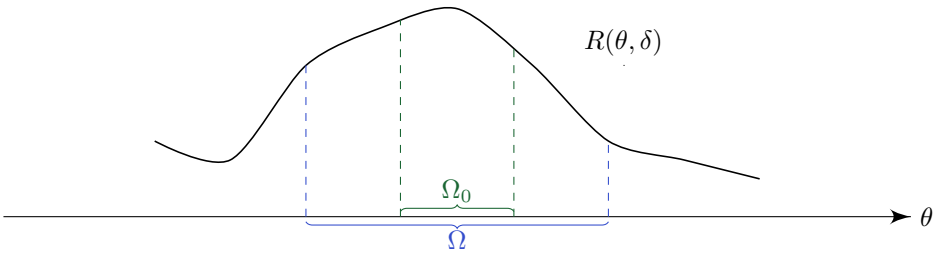
By Theorem 2.4.1 (ii), $\bar{X}$ is minimax (Not necessarily unique).

Proposition 2.4.1. (Most difficult models): Consider a submodel $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega_0\} \subseteq \mathcal{P} = \{P_\theta : \theta \in \Omega\}$, where $\Omega_0 \subseteq \Omega$. If

(i) $\delta(X)$ is a minimax estimator of $g(\theta)$ over Ω_0,

(ii) $\sup_{\theta\in\Omega_0} R(\theta, \delta) = \sup_{\theta\in\Omega} R(\theta, \delta)$.

Then $\delta(X)$ is minimax over Ω.



Proof. If δ' satisfied

$$\sup_{\theta\in\Omega} R(\theta, \delta') < \sup_{\theta\in\Omega} R(\theta, \delta) = \sup_{\theta\in\Omega_0} R(\theta, \delta),$$

then

$$\sup_{\theta\in\Omega_0} R(\theta, \delta') \le \sup_{\theta\in\Omega} R(\theta, \delta') < \sup_{\theta\in\Omega} R(\theta, \delta) = \sup_{\theta\in\Omega_0} R(\theta, \delta).$$

So δ is not a minimax on Ω_0, it's a contradiction. □

Example 2.4.3. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, $\sigma^2 \leq 10$ but any else unkown. What is the minimax estimator for μ over this class under squared-error loss?

Initiation: $\sigma^2 = 10$ provides the most difficult examples $\mathcal{P}_0 = \{N(\mu, 10), \mu \in \mathbb{R}\}$ which fixes $\sigma^2 = 10$. By Example 2.4.2 and (2.4.24), $\bar{X}$ is minimax over $\mathcal{P}_0$. Then the risk for $N(\mu, \sigma^2)$ is

$$R((\mu, \sigma^2), \bar{X}) = \mathbb{E}[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n}. \tag{2.4.27}$$

This is maximum when $\sigma^2 = 10$. So $\bar{X}$ remains minimax in the original model.

Example 2.4.4. Suppose $X_1, X_2, \cdots, X_n$ i.i.d. from some distribution with mean μ and unknown variance $\sigma^2 \leq 10$. What is the minimax estimator for μ over this class under squared-error loss?

Consider $\mathcal{P}_0 = \{N(\mu, 10), \mu \in \mathbb{R}\}$, $\bar{X}$ is minimax over $\mathcal{P}_0$.
For any such distribution F with mean μ and unknown variance σ^2, the risk is

$$R(F, \bar{X}) = \mathbb{E}[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n}. \tag{2.4.28}$$

It is still maximized on $\mathcal{P}_0$, so $\bar{X}$ remains minimax in the original model.

Example 2.4.5. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$, F supported on $[0, 1]$ with mean μ_F and variance ${\sigma_F}^2$. What is the minimax estimator for μ over this class under squared-error loss?

Guess: Consider the extreme case where the submodel is $Bernoulli(\mu_F)$. By Example 2.4.1 and (2.4.20), $\delta = \frac{\frac{\sqrt{n}}{2} + n\bar{X}}{\sqrt{n}+n}$ is a minimax estimator with constant risk $\frac{1}{4(1+\sqrt{n})^2}$.
Consider any such distribution F with mean μ_F and variance ${\sigma_F}^2$. The risk is

$$\begin{aligned} R(F, \delta) &= \mathbb{E}_F[(\delta(X) - \mu_F)^2] \\ &= (\mathbb{E}_F \delta(X) - \mu_F)^2 + Var_F[\delta(X)] \\ &= \left(\frac{\frac{\sqrt{n}}{2} + n\mu_F}{\sqrt{n}+n} - \mu_F\right)^2 + \left(\frac{n}{n+\sqrt{n}}\right)^2 \cdot \frac{{\sigma_F}^2}{n} \\ &= \frac{1}{(1+\sqrt{n})^2}\left(\left(\frac{1}{2} - \mu_F\right)^2 + {\sigma_F}^2\right). \end{aligned} \tag{2.4.29}$$

Note that $X^2 \leq X$ on $[0, 1]$,

$${\sigma_F}^2 = Var_F(X) = \mathbb{E}_F[X^2] - {\mu_F}^2 \leq \mathbb{E}_F[X] - {\mu_F}^2 = \mu_F - {\mu_F}^2$$

Then

$$R(F, \delta) \leq \frac{1}{(1+\sqrt{n})^2}\left(\left(\frac{1}{2} - \mu_F\right)^2 + \mu_F - {\mu_F}^2\right) = \frac{1}{4(1+\sqrt{n})^2}. \tag{2.4.30}$$

The worst-case risk of δ is achieved on the submodel $Bernoulli(\mu_F)$. So $\bar{X}$ is a minimax estimator over the class of all such distributions F.

Remark 2.4.3. An estimator δ is minimax if

$$\bar{r}(\delta) = \inf_{\delta'} \bar{r}(\delta') = \inf_{\delta'} \sup_{\theta \in \Omega} R(\theta, \delta') = \inf_{\delta'} \sup_{\Lambda} r(\Lambda, \delta'), \tag{2.4.31}$$

where $\sup_{\Lambda}$ is over all prior distributions Λ on Ω, and equality holds because

(i) For any prior Λ

$$r(\Lambda, \delta') \leq \sup_{\theta} R(\theta, \delta') \implies \inf_{\delta'} \sup_{\Lambda} r(\Lambda, \delta') \leq \inf_{\delta'} \sup_{\theta \in \Omega} R(\theta, \delta') \tag{2.4.32}$$

(ii) The class of all priors contains the point masses of any single parameter, so

$$\begin{aligned} &\sup_{\Lambda} r(\Lambda, \delta') \geq \sup_{\text{point mass priors } \lambda} r(\Lambda, \delta') = \sup_{\theta} R(\theta, \delta') \\ &\implies \inf_{\delta'} \sup_{\Lambda} r(\Lambda, \delta') \geq \inf_{\delta'} \sup_{\theta \in \Omega} R(\theta, \delta') \end{aligned} \tag{2.4.33}$$

Remark 2.4.1 and (2.4.6) is saying

$$\inf_{\delta} \sup_{\Lambda} r(\Lambda, \delta) \geq \sup_{\Lambda} \inf_{\delta} r(\Lambda, \delta) \tag{2.4.34}$$

Interpretation: Two-player game.

- I pick estimator δ, to minimize $r(\Lambda, \delta)$.
- Nature picks Λ, to maximize $r(\Lambda, \delta)$.

For convex loss function, with regularity condition, we expect (2.4.34) to hold with equality

$$\inf_{\delta} \sup_{\Lambda} r(\Lambda, \delta) = \sup_{\Lambda} \inf_{\delta} r(\Lambda, \delta). \tag{2.4.35}$$

because

- If loss is convex in δ, then $r(\Lambda, \delta)$ is also convex in δ.
- The space of all prior distributions Λ is a convex space.
- $r(\Lambda, \delta) = \int R(\theta, \delta) \mathrm{d}\Lambda(\theta)$ is linear in Λ.

The following theorem is a concrete result.

Theorem 2.4.2 (Kneser-Kuhn Theorem)**.** Let $f : K \times L \to \mathbb{R}$ where K, L are convex and L is compact. $f(x, y)$ is convex in x for all fixed y and concave and upper-semicontinuous in y for all fixed x. Then

$$\inf_{x \in K} \sup_{y \in L} f(x, y) = \sup_{y \in L} \inf_{x \in K} f(x, y). \tag{2.4.36}$$

We can show (2.4.35) holds if the paramter space Ω is compact, loss is convex, and any estimator δ where $R(\theta, \delta) < \infty \ \forall \ \theta \in \Omega$ has continuous risk over Ω.

2.5 Admissibility

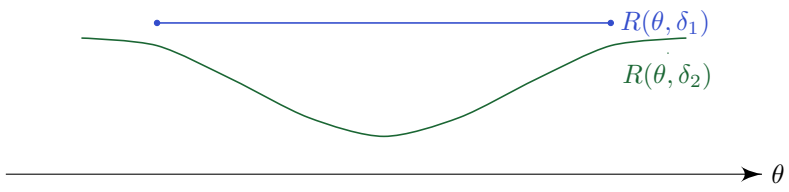
Recall: In Definition 2.1.3, for the estimators $\delta_1,\ \delta_2$ if

(i) $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ at all $\theta \in \Omega$.

(ii) $R(\theta, \delta_1) < R(\theta, \delta_2)$ at some $\theta \in \Omega$.

Then, δ_1 dominates δ_2, and δ_2 is inadmissable. If no estimator dominates δ_2, then δ_2 is admissable.

Note: δ is minimax doesn't imply δ is admissible.



It's possible that both δ_1 and δ_2 are minimax, but δ_1 here is inadmissible.

Proposition 2.5.1. If $\delta(X)$ is the unique Bayes estimator for any prior Λ, then it is admissible.

Proof. Suppose δ' dominates δ. Then

$$R(\theta, \delta') \leq R(\theta, \delta) \quad \forall\ \theta \in \Omega \implies \int R(\theta, \delta')\mathrm{d}\Lambda(\theta) \leq \int R(\theta, \delta)\mathrm{d}\Lambda(\theta). \quad (2.5.1)$$

Then we must have $\delta' = \delta$ because δ is the unique Bayes estimator. □

Example 2.5.1. $\delta(X) = \frac{\frac{\sqrt{n}}{2} + n\bar{X}}{\sqrt{n}+n}$ is admissble in the $Bernoulli(\theta)$ model.

Question 2.5.1. How to show $\bar{X}$ is admissible for estimating θ in $N(\theta, 1)$ model?

Answer 2.5.1. Blyth's method: use a sequence of prior distributions again.

Theorem 2.5.1 (Blyth's method)**.** Suppose $\delta(X)$ satisfies $R(\theta, \delta) < \infty$ for all $\theta \in \Omega$ and that any estimator δ' satisfying this condition has continuous risk in θ. Suppose Ω is open. If there exists a sequence of prior distributions $\Lambda_1, \Lambda_2, \cdots, \Lambda_i$ such that

(i) $r(\Lambda_i, \delta) < \infty$ for all i.

(ii) $\lim\limits_{i\to\infty} \frac{r(\Lambda_i,\delta) - B(\Lambda_i)}{\Lambda_i(\Omega_0)} = 0$ for every given subset $\Omega_0 \subseteq \Omega$.

Then δ is admissble.

Proof. Suppose that δ is dominated by δ', then

$$R(\theta, \delta') \leq R(\theta, \delta) \text{ for all } \theta \in \Omega,$$
$$R(\theta_0, \delta') < R(\theta_0, \delta) \text{ for some } \theta_0 \in \Omega.$$

Since Ω is open, and $R(\theta,\delta), R(\theta,\delta')$ are continuous in $\theta=\theta_0$, there exists an open neighborhood Ω_0 of θ_0 and $\epsilon>0$ where

$$R(\theta,\delta') \le R(\theta,\delta) - \epsilon \text{ for all } \theta \in \Omega_0.$$

Then

$$\begin{aligned} r(\Lambda_i,\delta) - r(\Lambda_i,\delta') &= \int R(\theta,\delta)\mathrm{d}\Lambda(\theta) - \int R(\theta,\delta')\mathrm{d}\Lambda(\theta) \\ &= \int (R(\theta,\delta) - R(\theta,\delta'))\mathrm{d}\Lambda(\theta) \\ &\ge \epsilon\Lambda_i(\Omega_0). \end{aligned}$$

So

$$\frac{r(\Lambda_i,\delta) - B(\Lambda_i)}{\Lambda_i(\Omega_0)} \ge \frac{r(\Lambda_i,\delta) - r(\Lambda_i,\delta')}{\Lambda_i(\Omega_0)} \ge \epsilon \text{ for all } i \tag{2.5.2}$$

This contradicts assumption (ii) in Blyth's method, so δ is admissible. □

Remark 2.5.1. In 1-parameter exponential family and under squared-error loss,

$$\theta \to R(\theta,\delta) = \int (\theta - \delta(x))^2 e^{\theta T(x) - A(\theta)} h(x)\mathrm{d}x$$

is continuous and differentiable. We can show this by applying Dominated Convergence Theorem and differentiating under the integral sign.

Example 2.5.2. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta,\sigma^2)$, σ^2 known. Let $\delta(X) = \bar{X}$. This is a 1-parameter exponential family, and if we consider the squared-error loss, by Remark 2.5.1, conditions of Blyth's method are satisfied. Take conjugate prior $\Lambda = N(0,\tau^2)$. By Example 2.4.2,

$$B(\Lambda) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \quad B(\Lambda) \to \frac{\sigma^2}{n} \text{ as } \tau^2 \to \infty.$$

$r(\Lambda,\bar{X}) = \dfrac{\sigma^2}{n}$. So $\bar{X}$ is minimax.

To apply Blyth's method

- $r(\Lambda,\bar{X}) - B(\Lambda) = \frac{\sigma^2}{n} - \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} = \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$.
- Let Ω_0 be an open set. This contains an interval (a,b).

$$\begin{aligned} \Lambda(a,b) &= \mathbb{P}_{\Theta\in\Omega}[\Theta \in (a,b)] = \mathbb{P}_{Z\sim N(0,1)}[z \in (\frac{a}{\tau},\frac{b}{\tau})] \\ &= \int_{\frac{a}{\tau}}^{\frac{b}{\tau}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\mathrm{d}z > \frac{b-a}{\tau}\cdot\frac{1}{\sqrt{2\pi}}(1-\epsilon) \equiv \frac{c}{\tau} \\ &\quad \text{for all large enough } \tau. \end{aligned}$$

So

$$\frac{r(\Lambda,\bar{X}) - B(\Lambda)}{\Lambda(a,b)} \le \frac{\frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}}{\frac{c}{\tau}} = \frac{1}{c(\frac{n}{\sigma^2}+\frac{1}{\tau^2})\cdot\tau} \to 0 \text{ as } \tau^2 \to \infty. \tag{2.5.3}$$

So $\bar{X}$ is admissible.

Remark 2.5.2. This argument works only in dimension $d=1$. If we consider arbitrary dimension, then $r(\Lambda,\bar{X}) - B(\Lambda) \propto \frac{1}{\tau^2}$ but $\Lambda(\Omega_0) \propto \frac{1}{\tau^d}$ for any field Ω_0.

2.6 Shrinkage Estimation

Setup: Suppose $X \sim N(\theta, \sigma^2 I) \in \mathbb{R}^d,\ \theta = (\theta_1, \theta_2, \cdots, \theta_d),\ \sigma^2$ known. We are going to estimate θ under Loss $L(\theta, \delta) = \sum_{i=1}^{d} (\theta_i - \delta_i)^2 = \|\theta - \delta\|^2$. We can generalize to $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2 I) \in \mathbb{R}^d$ by reducing to $\bar{X} \overset{\text{i.i.d.}}{\sim} N(\theta, \frac{\sigma^2}{n} I) \in \mathbb{R}^d$.

Unbiased analysis: $\delta(X) = X$ is an unbiased estimator.The risk is

$$R(\theta, \delta) = \mathbb{E}_\theta[\|\theta - X\|^2] = d\sigma^2. \tag{2.6.1}$$

Bayesian analysis: Suppose $\Theta \sim N(0, \tau^2 I) \equiv \Lambda$. Then $\theta_1, \theta_2, \cdots, \theta_d$ are independent, $X_1, X_2, \cdots, X_d$ are independent. So

$$\pi(\theta) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\theta_i^2}{2\tau^2}}, \quad \pi(x|\theta) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \theta_i)^2}{2\sigma^2}}.$$

$$\pi(\theta|x) \propto \prod_{i=1}^{d} e^{-\frac{\theta_i^2}{2\tau^2} - \frac{(x_i - \theta_i)^2}{2\sigma^2}} \qquad \theta_1, \theta_2, \cdots, \theta_d \text{ remain independent given } x.$$

$$\Theta | X \sim N\left(\frac{\frac{1}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} X, \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} \right).$$

The Bayes estimator is

$$\delta^{\text{Bayes}}(X) = cX, \quad c = \frac{\frac{1}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} = \frac{\tau^2}{\sigma^2 + \tau^2} < 1. \tag{2.6.2}$$

This is a *shrinkage estimator*.

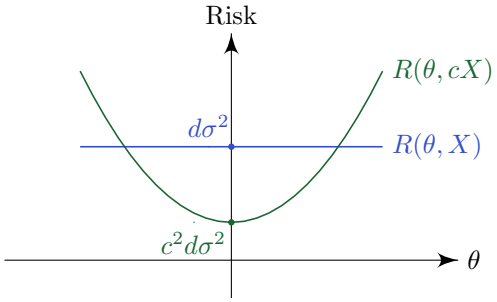
Minimax analysis: The Bayes risk of $\Lambda = N(0, \tau^2 I)$ is

$$B(\Lambda) = d \times B(\Lambda_i) = d \cdot \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} \quad \text{as } \tau^2 \to \infty,\ B(\Lambda) = d\sigma^2. \tag{2.6.3}$$

Here, $\Lambda_i = N(0, \sigma^2)$ is the prior distribution for θ_i. This certifies that $\delta(X) = X$ is a minimax estimator.

Risk for shrinkage estimation: For the shrinkage estimator $\delta(X) = cX$, the risk function is

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\|\theta - cX\|^2] \\ &= \mathbb{E}_\theta[\|\theta - c\theta + c\theta - cX\|^2] \\ &= \|\theta - c\theta\|^2 + \mathbb{E}_\theta[\|c\theta - cX\|^2] \\ &= (1 - c)^2 \|\theta\|^2 + c^2 \cdot d\sigma^2 \\ &= \text{Bias}^2 + \text{Variance}. \end{aligned} \tag{2.6.4}$$



$R(\theta, cX)$ can be much smaller if c is small and θ really is near 0.

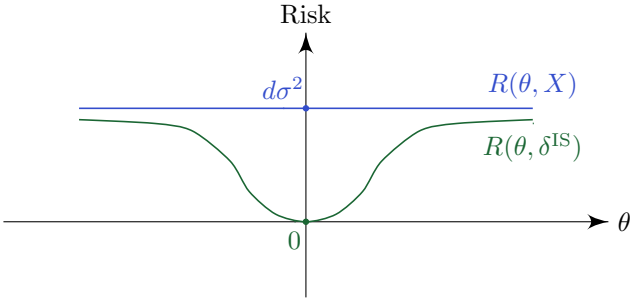
Ideal shrinkage estimation: Suppose we know $\|\theta\|$ and consider the estimator $\delta(X) = cX$, the ideal choice of c is

$$c^{\text{IS}} \equiv \underset{c}{\text{argmin}}[(1-c)^2\|\theta\|^2 + c^2 \cdot d\sigma^2] = \frac{\|\theta\|^2}{\|\theta\|^2 + d\sigma^2} = \frac{\frac{\|\theta\|^2}{d}}{\frac{\|\theta\|^2}{d} + \sigma^2}. \tag{2.6.5}$$

It corresponds to prior variance $\tau^2 = \frac{\|\theta\|^2}{d}$.
Let $\delta^{\text{IS}}(x) = c^{\text{IS}}x$. The ideal risk is

$$R(\theta, \delta^{\text{IS}}(x)) = (1-c^{\text{IS}})^2\|\theta\|^2 + (c^{\text{IS}})^2 \cdot d\sigma^2 = \frac{\|\theta\|^2 \cdot d\sigma^2}{\|\theta\|^2 + d\sigma^2}. \tag{2.6.6}$$

Note: $\delta^{\text{IS}}(x)$ is not a true estimator, it requires "oracle" knowledge of $\|\theta\|$. So $R(\theta, \delta^{\text{IS}})$ is not achievable.



Empirical Bayes analysis: We want to estimate $c = \frac{\tau^2}{\sigma^2+\tau^2} = 1 - \frac{\sigma^2}{\sigma^2+\tau^2}$ from the data. Since

$$X \sim N(\Theta, \sigma^2 I),\ \Theta \sim N(0, \tau^2 I).$$

Marginally,

$$X \sim N(0, (\sigma^2+\tau^2)I)$$
$$\|X\|^2 \sim (\sigma^2+\tau^2)\mathcal{X}_d^2$$

where $\mathcal{X}_d^2$ is a chi-squared distribution with d degrees of freedom.

Fact: If $Y \sim \mathcal{X}_d^2$, then $\mathbb{E}\left[\frac{1}{Y}\right] = \frac{1}{d-2}$ when $d > 2$.

Then

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\|X\|^2}\right] &= \frac{1}{\sigma^2+\tau^2}\cdot\frac{1}{d-2},\\ \mathbb{E}\left[1-\frac{d-2}{\|X\|^2}\cdot\sigma^2\right] &= 1-\frac{\sigma^2}{\sigma^2+\tau^2}=c.\end{aligned} \tag{2.6.7}$$

So $1-\frac{(d-2)\sigma^2}{\|X\|^2}$ is unbiased for c. Then an empirical Bayes estimator for θ is

$$\delta^{\mathrm{JS}}(X)=\left(1-\frac{(d-2)\sigma^2}{\|X\|^2}\right)X. \tag{2.6.8}$$

This is the James-Stein estimator.
Now we want to compute $R(\theta,\delta^{\mathrm{JS}})$.

Lemma 2.6.1 (Stein's Lemma)**.** Let $Z\sim N(0,1)$. Suppose $f:\mathbb{R}\to\mathbb{R}$ is absolutely continuous, differentiable and $\mathbb{E}[f'(Z)]<\infty$, $\mathbb{E}[Zf(Z)]<\infty$. Then

$$\mathbb{E}[Zf(Z)]=\mathbb{E}[f'(Z)]. \tag{2.6.9}$$

Proof. Let $\varphi(z)$ be the density function of Z,

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}.$$

Then

$$\varphi'(z)=-z\varphi(z).$$

So by integrating by parts,

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \int_{-\infty}^{\infty} f'(z)\varphi(z)\mathrm{d}z\\ &= f(z)\varphi(z)\Big|_{-\infty}^{\infty}-\int_{-\infty}^{\infty}f(z)\varphi'(z)\mathrm{d}z\\ &= f(z)\varphi(z)\Big|_{-\infty}^{\infty}+\int_{-\infty}^{\infty}zf(z)\varphi(z)\mathrm{d}z\end{aligned}$$

Since

$$\int_a^b f'(z)\varphi(z)\mathrm{d}z=f(b)\varphi(b)-f(a)\varphi(a)+\int_a^b zf(z)\varphi(z)\mathrm{d}z.$$

By assumptions $\mathbb{E}[|f'(Z)|]<\infty$ and $\mathbb{E}[|Zf(Z)|]<\infty$, we have

$$\begin{aligned}\lim_{a\to-\infty,\ b\to\infty}\int_a^b f'(z)\varphi(z)\mathrm{d}z &= \mathbb{E}[f'(Z)].\\ \lim_{a\to-\infty,\ b\to\infty}\int_a^b zf(z)\varphi(z)\mathrm{d}z &= \mathbb{E}[Zf(Z)].\end{aligned}$$

Thus $\lim_{b\to\infty}f(b)\varphi(b)$, $\lim_{a\to\infty}f(a)\varphi(a)$ exists. $\mathbb{E}[|Zf(Z)|]<\infty$ also implies these two limits must be 0. So

$$\mathbb{E}[Zf(Z)]=\mathbb{E}[f'(Z)]. \tag{2.6.10}$$

□

Corollary 2.6.1. Let $X \sim N(\theta, \sigma^2)$ and f as Stein's Lemma.

$$\mathbb{E}[(X-\theta)f(X)] = \sigma^2 \mathbb{E}[f'(X)]. \tag{2.6.11}$$

Proof. Let $X = \theta + \sigma Z$

$$\begin{aligned} \mathbb{E}[(X-\theta)f(X)] &= \sigma \mathbb{E}[Z f(\theta + \sigma Z)] \\ &= \sigma^2 \mathbb{E}[f'(\theta + \sigma Z)] \\ &= \sigma^2 \mathbb{E}[f'(X)]. \end{aligned} \tag{2.6.12}$$

□

Theorem 2.6.1 (Stein's unbiased risk estimate)**.** In the model $X \sim N(\theta, \sigma^2 I) \in \mathbb{R}^d$. Let $\delta(X)$ be an estimator of θ. Suppose $g(X) = \delta(X) - X$. For each $i \in \{1, 2, \cdots, d\}$ and $X_{-i} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_d) \in \mathbb{R}^{d-1}$. Let $g_i(X) = g(x_1, \cdots, x_{i-1}, X_i, x_{i+1}, \cdots, x_d)$. Suppose each $g_i : \mathbb{R} \to \mathbb{R}$ is differentiable and $\mathbb{E}[|X_i g_i(X)|] < \infty$, $\mathbb{E}\left[\left|\frac{\partial}{\partial X_i} g_i(X)\right|\right] < \infty$. Then

$$U(X) = d\sigma^2 + 2\sigma^2 \sum_{i=1}^{d} \frac{\partial}{\partial X_i} g_i(X) + \|g(X)\|^2 \tag{2.6.13}$$

is an unbiased estimator for squared-error loss $R(\theta, \delta) = \mathbb{E}_\theta[\|\theta - \delta(X)\|^2]$.

Proof. The risk is

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\|\delta(X) - \theta\|^2] \\ &= \mathbb{E}_\theta[\|g(X) + X - \theta\|^2] \\ &= \mathbb{E}_\theta[\|g(X)\|^2] + \mathbb{E}_\theta[\|X - \theta\|^2] + 2\mathbb{E}_\theta[(X-\theta)^T g(X)] \\ &= \mathbb{E}_\theta[\|g(X)\|^2] + d\sigma^2 + 2\mathbb{E}_\theta[(X-\theta)^T g(X)] \\ &= \mathbb{E}_\theta[\|g(X)\|^2] + d\sigma^2 + 2\sum_{i=1}^{d} \mathbb{E}_\theta[(X_i - \theta) g_i(X)]. \end{aligned} \tag{2.6.14}$$

By Stein's Lemma,

$$\begin{aligned} \mathbb{E}_\theta[(X_i - \theta) g_i(X)] &= \mathbb{E}_\theta[(X_i - \theta) g(x_1, \cdots, x_{i-1}, X_i, x_{i+1}, \cdots, x_d)] \\ &= \sigma^2 \mathbb{E}_\theta\left[\frac{\partial}{\partial X_i} g(x_1, \cdots, x_{i-1}, X_i, x_{i+1}, \cdots, x_d)\right] \\ &= \sigma^2 \mathbb{E}_\theta\left[\frac{\partial}{\partial X_i} g_i(X)\right] \end{aligned} \tag{2.6.15}$$

Therefore,

$$R(\theta, \delta) = \mathbb{E}_\theta\left[d\sigma^2 + 2\sigma^2 \sum_{i=1}^{d} \frac{\partial}{\partial X_i} g_i(X) + \|g(X)\|^2\right] = \mathbb{E}_\theta[U(X)]. \tag{2.6.16}$$

□

Remark 2.6.1. $U(X)$ is a true estimator and doesn't depend on θ. Given a family of estimators $\delta_\lambda(X)$ depending on a tuning parameter λ, we can actually compute the risk estimators $U_\lambda(X)$. And we choose λ which minimizes $U_\lambda(X)$. This is a common alternative to cross-validation for tuning λ.

Theorem 2.6.2. For $X \sim N(\theta, \sigma^2 I) \in \mathbb{R}^d$ and $\delta^{\text{JS}}(X) = \left(1 - \frac{(d-2)\sigma^2}{\|X\|^2}\right) X$.

$$R(\theta, \delta^{\text{JS}}(X)) \leq 2\sigma^2 + \frac{(d-2)\sigma^2 \cdot \|\theta\|^2}{(d-2)\sigma^2 + \|\theta\|^2} < d\sigma^2 \tag{2.6.17}$$

Proof. By Stein's unbiased risk estimate,

$$\begin{aligned} &R(\theta, \delta^{\text{JS}}) = \mathbb{E}_\theta[U(X)], \\ &U(X) = d\sigma^2 + 2\sigma^2 \sum_{i=1}^{d} \frac{\partial}{\partial X_i} g_i(X) + \|g(X)\|^2, \\ &g(X) = \delta^{\text{JS}}(X) - X = -\frac{(d-2)\sigma^2}{\|X\|^2} X, \\ &g_i(X) = -\frac{(d-2)\sigma^2}{\|X\|^2} X_i. \end{aligned} \tag{2.6.18}$$

– For $\|g(X)\|^2$

$$\|g(X)\|^2 = \left(\frac{(d-2)\sigma^2}{\|X\|^2}\right)^2 \cdot \|X\|^2 = \frac{(d-2)^2\sigma^4}{\|X\|^2}. \tag{2.6.19}$$

– For $\sum_{i=1}^{d} \frac{\partial}{\partial X_i} g_i(X)$

$$\begin{aligned} \sum_{i=1}^{d} \frac{\partial}{\partial X_i} g_i(X) &= \sum_{i=1}^{d} -\frac{\partial}{\partial X_i} \frac{(d-2)\sigma^2}{X_1^2 + X_2^2 + \cdots + X_n^2} X_i \\ &= \sum_{i=1}^{d} \left(\frac{2(d-2)\sigma^2}{(X_1^2 + X_2^2 + \cdots + X_n^2)^2} X_i^2 - \frac{(d-2)\sigma^2}{X_1^2 + X_2^2 + \cdots + X_n^2} \right) \\ &= \frac{2(d-2)\sigma^2}{\|X\|^2} - \frac{d(d-2)\sigma^2}{\|X\|^2} \\ &= -\frac{(d-2)^2\sigma^2}{\|X\|^2}. \end{aligned} \tag{2.6.20}$$

Therefore

$$\begin{aligned} &U(X) = d\sigma^2 - \frac{(d-2)^2\sigma^4}{\|X\|^2}. \\ &R(\theta, \delta^{\text{JS}}) = \mathbb{E}_\theta[U(X)] = d\sigma^2 - (d-2)^2\sigma^4 \mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right] < d\sigma^2. \end{aligned} \tag{2.6.21}$$

So for any $d > 2$, James-Stein estimator dominates $\delta(X) = X$.
Fact: Let $X \sim N(\theta, \sigma^2 I)$. Then $\frac{1}{\sigma^2}\|X\|^2 \sim \mathcal{X}_d^2\left(\frac{\|\theta\|^2}{\sigma^2}\right)$, a non-central chi-squared distribution with d degrees of freedom and $\frac{\|\theta\|^2}{\sigma^2}$ non-centrality. This is a Poisson mixture of central chi-squared distributions $\mathcal{X}_{d+2N}^2$ where $N \sim Poisson\left(\frac{\|\theta\|^2}{2\sigma^2}\right)$.

Then

$$\begin{aligned}
\mathbb{E}\left[\frac{\sigma^2}{\|X\|^2}\right] &= \mathbb{E}\left[\frac{1}{\mathcal{X}^2_{d+2N}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\mathcal{X}^2_{d+2N}}\middle| N\right]\right] \\
&= \mathbb{E}\left[\frac{1}{d+2N-2}\right] \\
&\geq \frac{1}{\mathbb{E}[d+2N-2]} \text{ By Jensen's Inequality for } x \to \frac{1}{x} \\
&= \frac{1}{d-2+\frac{\|\theta\|^2}{2\sigma^2}} \\
&= \frac{\sigma^2}{(d-2)\sigma^2+\|\theta\|^2}
\end{aligned} \tag{2.6.22}$$

Therefore,

$$R(\theta, \delta^{\mathrm{JS}}) \leq d\sigma^2 - \frac{(d-2)^2\sigma^4}{(d-2)\sigma^2+\|\theta\|^2} = 2\sigma^2 + \frac{(d-2)\sigma^2\cdot\|\theta\|^2}{(d-2)\sigma^2+\|\theta\|^2}. \tag{2.6.23}$$

Note that $R(\theta, \delta^{\mathrm{JS}}) \to d\sigma^2$ as $\|\theta\|^2 \to \infty$, while $R(\theta, \delta^{\mathrm{JS}}) = 2\sigma^2 \ll d\sigma^2$ at $\theta = 0$ if d is large. □

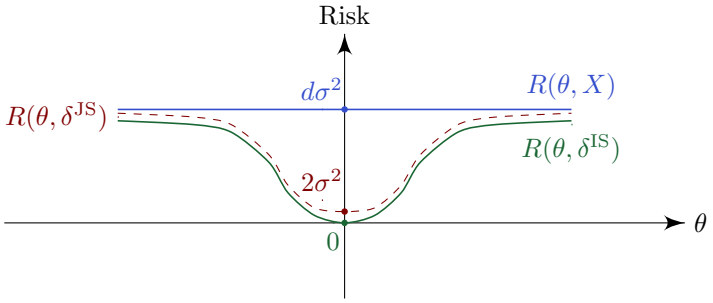
Remark 2.6.2. To compare $R(\theta, \delta^{\mathrm{JS}})$ with $R(\theta, \delta^{\mathrm{IS}})$, by (2.6.6)

$$R(\theta, \delta^{\mathrm{IS}}) = \frac{d\sigma^2\|\theta\|^2}{d\sigma^2+\|\theta\|^2} \tag{2.6.24}$$

Theorem 2.6.2 and (2.6.17) implies

$$R(\theta, \delta^{\mathrm{JS}}) \leq 2\sigma^2 + R(\theta, \delta^{\mathrm{IS}}). \tag{2.6.25}$$

So the increase in risk over the unachievable "oracle" risk is small: $2\sigma^2$. This is often called an *oracle inequality*.

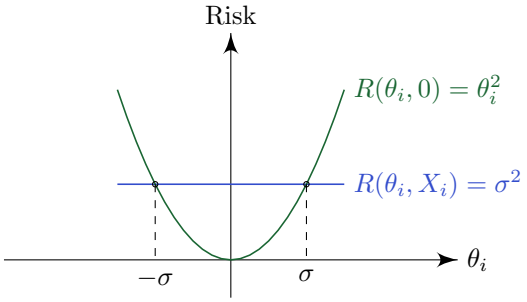


2.7 Sparsity and Thresholding

Setup: In James-Stein, we consider $X \sim N(\theta, \sigma^2) \in \mathbb{R}^d$, $\theta = (\theta_1, \theta_2, \cdots, \theta_d)$. If we believe $\|\theta\|^2$ is small, we can decrease risk by shrinking the estimate $\delta(X) = X$ towards 0. We can estimate how much to shrink using empirical Bayes.

Question 2.7.1. What if we believe only a few coordinates of θ are "large", and the rest are either 0 or "very small" (i.e., θ is sparse)?

Intuition: We estimate large coordinates θ_i by X_i and estimate small coordinates θ_i by 0. And 0 is a better estimate than X_i if $\theta_i^2 \le \sigma^2$.



Ideal thresholding estimator: Suppose an "oracle" tells us which coordinates θ_i^2 satifies $\theta_i^2 \ge \sigma^2$. Then we can consider the ideal thresholding estimator

$$\delta^{\mathrm{I}}(X)_i = \begin{cases} X_i & \text{if } \theta_i^2 \ge \sigma^2 \\ 0 & \text{if } \theta_i^2 < \sigma^2 \end{cases} \tag{2.7.1}$$

This would achieve the ideal risk under squared-error loss

$$R(\theta, \delta^{\mathrm{I}}) = \sum_{i=1}^{d} \min(\theta_i^2, \sigma^2) \tag{2.7.2}$$

Example 2.7.1. If θ is exactly sparse with k large coordinates and remaining coordinates 0, then

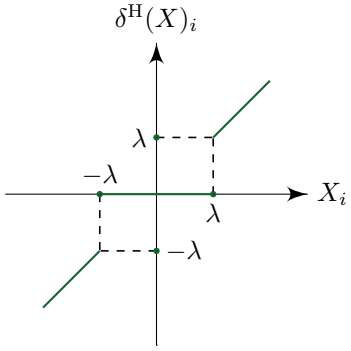
$$R(\theta, \delta^{\mathrm{I}}) = k\sigma^2 \ll d\sigma^2 \text{ when } k \ll d. \tag{2.7.3}$$

Note 2.7.1. δ^{I} is not an estimator, and $R(\theta, \delta^{\mathrm{I}})$ is unachievable without this oracle.

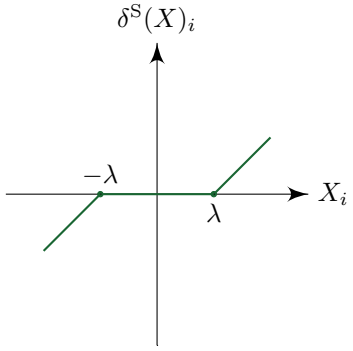
Thresholding estimators:

$$\delta^{\mathrm{H}}(X)_i = \begin{cases} X_i & \text{if } |X_i| \ge \lambda \\ 0 & \text{if } |X_i| < \lambda \end{cases} \tag{2.7.4}$$

$$\delta^{\mathrm{S}}(X)_i = \begin{cases} X_i - \lambda & \text{if } X_i \ge \lambda \\ 0 & \text{if } |X_i| < \lambda \\ X_i + \lambda & \text{if } X_i \le -\lambda \end{cases} \tag{2.7.5}$$



Hard-thresholding



Soft-thresholding

Equivalently:

- $\delta^{\mathrm{H}}(X) = \underset{\delta}{\operatorname{argmin}} \|X-\delta\|_2^2 + \lambda^2\|\delta\|_0, \quad \|\delta\|_0 = \sum_{i=1}^{d} \mathbb{1}\{\delta_i \neq 0\}.$
- $\delta^{\mathrm{S}}(X) = \underset{\delta}{\operatorname{argmin}} \|X-\delta\|_2^2 + 2\lambda\|\delta\|_1, \quad \|\delta\|_1 = \sum_{i=1}^{d} |\delta_i|.$

Proposition 2.7.1 (Mill's ratio)**.** Suppose $Z \sim N(0,1)$ and $t > 0$, then

$$\mathbb{P}[Z \geq t] \leq \frac{1}{t}\varphi(t), \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \tag{2.7.6}$$

Proof. Note $\varphi'(t) = -t\varphi(t)$. Then

$$\begin{aligned} \mathbb{P}[Z \geq t] &= \int_t^\infty \varphi(s)\mathrm{d}s \\ &\leq \int_t^\infty \frac{s}{t}\varphi(s)\mathrm{d}s \\ &= \frac{1}{t}\int_t^\infty [-\varphi'(s)]\mathrm{d}s \\ &= -\frac{1}{t}[\varphi]_t^\infty = \frac{1}{t}\varphi(t). \end{aligned} \tag{2.7.7}$$

□

Corollary 2.7.1. In hard-thresholding (2.4.7), set $\lambda = \sigma\sqrt{2\log d}$. Then

$$\mathbb{P}[|X_i| \geq \lambda \text{ for some } i \text{ where } \theta_i = 0] \leq \frac{1}{\sqrt{\pi \log d}}. \tag{2.7.8}$$

Proof.

$$\begin{aligned} \mathbb{P}[|X_i| \geq \lambda \text{ for some } i \text{ where } \theta_i = 0] &\leq \sum_{i:\theta_i=0} \mathbb{P}[|X_i| \geq \lambda] \\ &\leq d \cdot \mathbb{P}[|Z| \geq \frac{\lambda}{\sigma}] \\ &\leq d \cdot 2\mathbb{P}[Z \geq \frac{\lambda}{\sigma}] \\ &= 2d \cdot \mathbb{P}[Z \geq \sqrt{2\log d}] \quad (Z \sim N(0,1)). \end{aligned}$$

By Mill's ratio,

$$\begin{aligned} \mathbb{P}[Z \geq \sqrt{2\log d}] &\leq \frac{1}{\sqrt{2\log d}} \cdot \varphi(\sqrt{2\log d}) \\ &\leq \frac{1}{\sqrt{2\log d}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{2\log d}{2}} \\ &= \frac{1}{2d} \cdot \frac{1}{\sqrt{\pi \log d}}. \end{aligned} \tag{2.7.9}$$

□

Remark 2.7.1. $\lambda = \sigma\sqrt{2\log d}$ is called a "universal threshold" and it ensures no false discoveries with probability $\to 1$ as $d \to \infty$.

Lemma 2.7.1. Consider the univariate problem $X \sim N(\theta, 1) \in \mathbb{R}$, and let $\delta^{\mathrm{S}}(X)$ be soft-thresholding at $\lambda > 1$. Let $R(\theta, \delta^{\mathrm{S}}) = \mathbb{E}_\theta[(\theta - \delta^{\mathrm{S}}(X))^2]$. Then

(i) If $\theta = 0$, $R(0, \delta^{\mathrm{S}}) \le e^{-\frac{\lambda^2}{2}}$.

(ii) For all $\theta \in \mathbb{R}$, $R(\theta, \delta^{\mathrm{S}}) \le R(0, \delta^{\mathrm{S}}) + \theta^2$.

(iii) For all $\theta \in \mathbb{R}$, $R(\theta, \delta^{\mathrm{S}}) \le 1 + \lambda^2$.

Proof. By (2.7.5),

$$\delta^{\mathrm{S}}(X)_i = \begin{cases} X_i - \lambda & \text{if } X_i \ge \lambda \\ 0 & \text{if } |X_i| < \lambda \\ X_i + \lambda & \text{if } X_i \le -\lambda \end{cases} \tag{2.7.10}$$

(i) If $\theta = 0$,

$$\begin{aligned} R(0, \delta^{\mathrm{S}}) &= \mathbb{E}_0[\delta^{\mathrm{S}}(X)^2] \\ &= \int_\lambda^\infty (x-\lambda)^2 \varphi(x)\mathrm{d}x + \int_{-\infty}^{-\lambda} (x+\lambda)^2 \varphi(x)\mathrm{d}x \\ &= 2\int_\lambda^\infty (x-\lambda)^2 \varphi(x)\mathrm{d}x \\ &= 2\left(\int_\lambda^\infty x^2\varphi(x)\mathrm{d}x - 2\lambda\int_\lambda^\infty x\varphi(x)\mathrm{d}x + \lambda^2\int_\lambda^\infty \varphi(x)\mathrm{d}x\right). \end{aligned}$$

Suppose $\Phi(x) = \mathbb{P}[Z \le x]$ and $\tilde{\Phi}(x) = \mathbb{P}[Z \ge x]$ for $Z \sim N(0,1)$. Apply

- $\int_\lambda^\infty \varphi(x)\mathrm{d}x = \tilde{\Phi}(\lambda)$.
- $\int_\lambda^\infty x\varphi(x)\mathrm{d}x = [-\varphi(x)]_\lambda^\infty = \varphi(\lambda)$.
- $\int_\lambda^\infty x^2\varphi(x)\mathrm{d}x = \int_\lambda^\infty (-x)(-x\varphi(x))\mathrm{d}x = [-x\varphi(x)]_\lambda^\infty + \int_\lambda^\infty \varphi(x)\mathrm{d}x$

  $\int_\lambda^\infty x^2\varphi(x)\mathrm{d}x = \lambda\varphi(\lambda) + \tilde{\Phi}(\lambda)$.

Then by Mill's ratio,

$$\begin{aligned} R(0, \delta^{\mathrm{S}}) &= 2(\lambda\varphi(\lambda) + \tilde{\Phi}(\lambda) - 2\lambda\varphi(\lambda) + \lambda^2\tilde{\Phi}(\lambda)) \\ &= 2(\lambda^2+1)\tilde{\Phi}(\lambda) - 2\lambda\varphi(\lambda) \\ &\le 2(\lambda^2+1)\cdot\frac{1}{\lambda}\varphi(\lambda) - 2\lambda\varphi(\lambda) \\ &= \frac{2}{\lambda}\varphi(\lambda) \\ &< e^{-\frac{\lambda^2}{2}} \quad \text{for } \lambda > 1. \end{aligned} \tag{2.7.11}$$

(ii) For all $\theta \in \mathbb{R}$,

$$\begin{aligned} R(\theta, \delta^{\mathrm{S}}) &= \mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - \theta)^2] \\ &= \theta^2 \int_{-\lambda}^{\lambda} \varphi(x-\theta)\mathrm{d}x + \int_{\lambda}^{\infty} (x-\lambda-\theta)^2 \varphi(x-\theta)\mathrm{d}x \\ &+ \int_{-\infty}^{-\lambda} (x+\lambda-\theta)^2 \varphi(x-\theta)\mathrm{d}x \\ &= \theta^2 \int_{-\lambda-\theta}^{\lambda-\theta} \varphi(z)\mathrm{d}z + \int_{\lambda-\theta}^{\infty} (z-\lambda)^2 \varphi(z)\mathrm{d}z \\ &+ \int_{-\infty}^{-\lambda-\theta} (z+\lambda)^2 \varphi(z)\mathrm{d}z \end{aligned}$$

Apply $\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(z)\mathrm{d}z = f(x), \quad \frac{\mathrm{d}}{\mathrm{d}x}\int_x^a f(z)\mathrm{d}z = -f(x)$.

$$\begin{aligned} \frac{\partial}{\partial\theta} R(\theta, \delta^{\mathrm{S}}) &= 2\theta \int_{-\lambda-\theta}^{\lambda-\theta} \varphi(z)\mathrm{d}z + \theta^2(-\varphi(\lambda-\theta) + \varphi(-\lambda-\theta)) \\ &+ (\lambda-\theta-\lambda)^2\varphi(\lambda-\theta) - (-\lambda-\theta+\lambda)^2\varphi(-\lambda-\theta) \\ &= 2\theta \int_{-\lambda-\theta}^{\lambda-\theta} \varphi(z)\mathrm{d}z \\ &= 2\theta(\Phi(\lambda-\theta) - \Phi(-\lambda-\theta)) \le 2\theta \quad \text{for } \theta > 0. \end{aligned}$$

Then

$$\begin{aligned} R(\theta, \delta^{\mathrm{S}}) - R(0, \delta^{\mathrm{S}}) &= \int_0^\theta \frac{\partial}{\partial t} R(t, \delta^{\mathrm{S}})\mathrm{d}t \\ &\le \int_0^\theta 2t\mathrm{d}t = \theta^2 \quad \text{for } \theta > 0. \end{aligned} \tag{2.7.12}$$

For $\theta < 0$

$$\begin{aligned} R(\theta, \delta^{\mathrm{S}}) &= \mathbb{E}_\theta[(\theta - \delta^{\mathrm{S}}(X))^2] \\ &= \mathbb{E}_{-\theta}[(\theta - \delta^{\mathrm{S}}(-X))^2] \\ &= \mathbb{E}_{-\theta}[(\theta + \delta^{\mathrm{S}}(X))^2] \\ &= R(-\theta, \delta^{\mathrm{S}}(X)). \end{aligned} \tag{2.7.13}$$

Therefore

$$R(\theta, \delta^{\mathrm{S}}) \le R(0, \delta^{\mathrm{S}}) + \theta^2. \tag{2.7.14}$$

(iii) For all $\theta \in \mathbb{R}$,

$$\begin{aligned} R(\theta, \delta^{\mathrm{S}}) &= \mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - \theta)^2] \\ &= \mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - X + X - \theta)^2] \\ &= \mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - X)^2] + \mathbb{E}_\theta[(X-\theta)^2] + 2\mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - X)(X-\theta)] \\ &\le \lambda^2 + 1 + 2\mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - X)(X-\theta)] \end{aligned}$$

Note: If f is decreasing and g is increasing, then

$$Cov(f(X), g(X)) = \mathbb{E}[f(X)g(X)] - \mathbb{E}[f(X)]\mathbb{E}[g(X)] \le 0.$$

Proof. Let Y be an independent copy of X. Then

$$(f(X) - f(Y))(g(X) - g(Y)) \leq 0.$$

So

$$\begin{aligned} 0 &\geq \mathbb{E}[(f(X) - f(Y))(g(X) - g(Y))] \\ &= \mathbb{E}[f(X)g(X)] + \mathbb{E}[f(Y)g(Y)] - \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(Y)g(X)] \\ &= 2\mathbb{E}[f(X)g(X)] - 2\mathbb{E}[f(X)]\mathbb{E}[g(X)]. \end{aligned}$$

Apply this to $f(X) = \delta^{\mathrm{S}}(X) - X$ (decreasing), $g(X) = X - \theta$ (increasing). Then

$$\mathbb{E}_\theta[(\delta^{\mathrm{S}}(X) - X)(X - \theta)] \leq \mathbb{E}_\theta[\delta^{\mathrm{S}}(X) - X] \cdot \mathbb{E}_\theta[X - \theta] = 0$$

□

So

$$R(\theta, \delta^{\mathrm{S}}) \leq 1 + \lambda^2. \tag{2.7.15}$$

□

Corollary 2.7.2. For $X \sim N(\theta, \sigma^2)$, letting $\delta^{\mathrm{S}}(X)$ be soft-thresholding at $\lambda > \sigma$. Then

(i) If $\theta = 0$, $R(0, \delta^{\mathrm{S}}) \leq \sigma^2 e^{-\frac{\lambda^2}{2\sigma^2}}$.

(ii) For all $\theta \in \mathbb{R}$, $R(\theta, \delta^{\mathrm{S}}) \leq R(0, \delta^{\mathrm{S}}) + \theta^2$.

(iii) For all $\theta \in \mathbb{R}$, $R(\theta, \delta^{\mathrm{S}}) \leq \sigma^2 + \lambda^2$.

Proof. Note that $\frac{X}{\sigma} \sim N(\frac{\theta}{\sigma}, 1)$. Let $\tilde{\delta}^{\mathrm{S}}(X)$ be soft-thresholding at $\frac{\theta}{\sigma}$. Then

$\delta^{\mathrm{S}}(X) = \sigma \cdot \tilde{\delta}^{\mathrm{S}}(\frac{X}{\sigma})$.

$$R(\theta, \delta^{\mathrm{S}}(X)) = \mathbb{E}_\theta\left[\left(\theta - \sigma \cdot \tilde{\delta}^{\mathrm{S}}\left(\frac{X}{\sigma}\right)\right)^2\right] = \sigma^2 \mathbb{E}_\theta\left[\left(\frac{\theta}{\sigma} - \tilde{\delta}^{\mathrm{S}}\left(\frac{X}{\sigma}\right)\right)^2\right] = \sigma^2 R\left(\frac{\theta}{\sigma}, \tilde{\delta}^{\mathrm{S}}\left(\frac{X}{\sigma}\right)\right).$$

$R\left(\frac{\theta}{\sigma}, \tilde{\delta}^{\mathrm{S}}\left(\frac{X}{\sigma}\right)\right) \equiv \tilde{R}(\xi, \tilde{\delta}^{\mathrm{S}}(Z)) \equiv \tilde{R}(\xi, \tilde{\delta}^{\mathrm{S}})$ is the risk in the model $Z \sim N(\xi, 1)$.

So by Lemma 2.7.1

(i) If $\theta = 0$, $R(0, \delta^{\mathrm{S}}) = \sigma^2 \tilde{R}(0, \tilde{\delta}^{\mathrm{S}}) \leq \sigma^2 e^{-\frac{\lambda^2}{2\sigma^2}}$.

(ii) For all $\theta \in \mathbb{R}$, $R(\theta, \delta^{\mathrm{S}}) = \sigma^2 \tilde{R}(\frac{\theta}{\sigma}, \tilde{\delta}^{\mathrm{S}}) \leq \sigma^2(\tilde{R}(0, \tilde{\delta}^{\mathrm{S}}) + (\frac{\theta}{\sigma})^2) \leq R(0, \delta^{\mathrm{S}}) + \theta^2$.

(iii) For all $\theta \in \mathbb{R}$, $R(\theta, \delta^{\mathrm{S}}) = \sigma^2 \tilde{R}(\frac{\theta}{\sigma}, \tilde{\delta}^{\mathrm{S}}) \leq \sigma^2(1 + (\frac{\lambda}{\sigma})^2) \leq \sigma^2 + \lambda^2$.

□

Theorem 2.7.1. Let δ^{S} be the soft-thresholding at $\lambda = \sigma\sqrt{2\log d}$. Let $R(\theta, \delta^{\mathrm{I}}) = \sum_{i=1}^{d} \min(\theta_i^2, \sigma^2)$ be the ideal risk. Then the squared-error risk of δ^{S} satisfies

$$R(\theta, \delta^{\mathrm{S}}) \leq \sigma^2 + (1 + 2\log d) R(\theta, \delta^{\mathrm{I}}). \tag{2.7.16}$$

Note 2.7.2. For $\delta^{\mathrm{H}}(X)$ the hard-thresholding estimator, more involved arguments show

$$R(\theta, \delta^{\mathrm{H}}) \le (2\log d + 1.2)(\sigma^2 + R(\theta, \delta^{\mathrm{I}})). \tag{2.7.17}$$

Proof. By Corollary 2.7.2,

$$\begin{aligned} R(\theta, \delta^{\mathrm{S}}) &= \mathbb{E}_\theta[\|\theta - \delta^{\mathrm{S}}\|^2] \\ &\le \sum_{i=1}^{d} \min(\sigma^2 \cdot e^{-\frac{\lambda^2}{2\sigma^2}} + \theta_i^2, \sigma^2 + \lambda^2). \end{aligned}$$

Take $\lambda = \sigma\sqrt{2\log d}$

$$\begin{aligned} R(\theta, \delta^{\mathrm{S}}) &\le \sum_{i=1}^{d} \min(\frac{\sigma^2}{d} + \theta_i^2, \sigma^2 + 2\sigma^2 \log d) \\ &\le \frac{\sigma^2}{d} + \sum_{i=1}^{d} \min(\theta_i^2, \sigma^2)(1 + 2\log d) \\ &= \sigma^2 + (1 + 2\log d) R(\theta, \delta^{\mathrm{I}}). \end{aligned} \tag{2.7.18}$$

□

Remark 2.7.2. This says

$$R(\theta, \delta^{\mathrm{S}}) \lesssim (2\log d) \sum_{i=1}^{d} \min(\theta_i^2, \sigma^2). \tag{2.7.19}$$

By Theorem 2.6.2 and (2.6.17), for James-Stein estimator

$$R(\theta, \delta^{\mathrm{JS}}(X)) \lesssim \frac{d\sigma^2 \cdot \|\theta\|^2}{d\sigma^2 + \|\theta\|^2}.$$

Since

$$\frac{1}{2}\min(a, b) \le \frac{ab}{a+b} \le \min(a, b).$$

Then

$$R(\theta, \delta^{\mathrm{JS}}(X)) \lesssim \frac{d\sigma^2 \cdot \|\theta\|^2}{d\sigma^2 + \|\theta\|^2} \in \left[\frac{1}{2}, 1\right] \cdot \min(\|\theta\|^2, d\sigma^2). \tag{2.7.20}$$

– Scenario 1: θ is exactly sparse with k large coordinates and remaining coordinates 0: $\theta = (c, \cdots, c, 0, \cdots, 0)$ where $c > \sigma$. Then

$$\begin{aligned} &(2\log d) \sum_{i=1}^{d} \min(\theta_i^2, \sigma^2) = (2\log d) \cdot k\sigma^2 \ll \min(\|\theta\|^2, d\sigma^2). \\ &\text{If } k \ll \frac{d}{\log d}, \quad c^2 \gg \sigma^2 \cdot \log d. \end{aligned} \tag{2.7.21}$$

– Scenario 2: θ is exactly dense with d small coordinates: $\theta = (c, c, \cdots, c)$. Then

$$(2\log d) \sum_{i=1}^{d} \min(\theta_i^2, \sigma^2) = (2\log d) \cdot d \cdot \min(c^2, \sigma^2) = (2\log d) \cdot \min(\|\theta\|^2, \sigma^2). \tag{2.7.22}$$

Factor $\sim \log d$ inflation over James Stein risk.

Remark 2.7.3. The extra $\log d$ factor over $R(\theta, \delta^{\mathrm{I}}(X))$ is necessary in a minimax sense:

$$\inf_{\delta} \sup_{\theta \in \mathbb{R}^d} \frac{R(\theta, \delta)}{\sigma^2 + \sum_i \min(\theta_i^2, \sigma^2)} \geq (2 \log d)(1 + o(1)). \tag{2.7.23}$$

As $d \to \infty$, we can show this by showing, for the loss

$$L(\theta, \delta) = \frac{\|\theta - \delta\|^2}{\sigma^2 + \sum_i \min(\theta_i^2, \sigma^2)}, \tag{2.7.24}$$

the Bayes risk

$$B(\Lambda) \geq (2 \log d)(1 + o(1)) \tag{2.7.25}$$

for a sparse prior Λ which sets $\theta_i \approx \sqrt{2 \log d} - c \cdot \log \log d$ for $\log d$ random coordinates i and $\theta_i = 0$ for all remaining coordinates (See Johnstone Chapter 8.5-8.6).

3 Hypothesis Testing

3.1 Hypothesis, Test, Size and Power

Setup: In Model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, our goal is to decide whether $\theta \in \Omega_0$ or $\theta \in \Omega_1$ for disjoint Ω_0, Ω_1 given data $X \sim P_\theta$.

- Null hypothesis $H_0 : \theta \in \Omega_0 \subset \Omega$.
- Alternative hypothesis $H_1 : \theta \in \Omega_1 \subset \Omega$.

Definition 3.1.1 (Test and rejection region)**.**

- A *test* $\phi : \mathcal{X} \to \{0, 1\}$. Value 1 means "reject H_0" and value 0 means "accept H_0".
- The *rejection region* is $\{x \in \mathcal{X} : \phi(x) = 1\}$.
- More generally, a *randomized test* $\phi : \mathcal{X} \to [0, 1]$ is a function giving each $x \in \mathcal{X}$ a probability of rejecting H_0, $\phi(x) \in [0, 1]$.

Note 3.1.1. The probability of rejecting H_0 is

$$\mathbb{E}_\theta[\phi(X) = 1] = \int \phi(x) \cdot p_\theta(x) \mathrm{d}x. \tag{3.1.1}$$

When ϕ is non-randomized, this is simply $\mathbb{P}_\theta[\phi(x) = 1]$.

Definition 3.1.2 (Size and power)**.**

- A *Type I error* occurs when we reject H_0, when in fact $\theta \in \Omega_0$.
- A *Type II error* occurs when we accept H_0, when in fact $\theta \in \Omega_1$.
- For $\theta \in \Omega_0$, the *size* of a test ϕ is $\sup_{\theta \in \Omega_0} \mathbb{E}_\theta[\phi(X)]$ (i.e., the maximum probability of making Type I error under any $\theta \in \Omega_0$).
- For $\theta \in \Omega_1$, the *power* of the test is $\beta(\theta) = \mathbb{E}_\theta[\phi(X)]$.

Remark 3.1.1. Consequences of Type I and Type II errors are asymmetric. And we want to control Type I error at some specified level (i.e., $\alpha = 0.05, 0.01, \cdots$), which is the *significance level* of the test while maximizing the power.

Example 3.1.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$. Test

$$H_0 : \theta = 0 \ (\Omega_0 = \{0\}) \text{ vs. } H_1 : \theta > 0 \ (\Omega_1 = \text{ positive real line}).$$

Let

$$\phi(X) = \begin{cases} 1 & \bar{X} > t \\ 0 & \bar{X} \le t \end{cases}, \text{ where } \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We want to choose t to ensure Type I error control at level α. The rejection region is $\{(X_1, X_2, \cdots, X_n) : \bar{X} > t\}$. The rejection probability for any θ is

$$\begin{aligned} \mathbb{P}_\theta[\phi(X) = 1] &= \mathbb{P}_\theta[\bar{X} > t] \\ &= \mathbb{P}_\theta[\sqrt{n}(\bar{X} - \theta) > \sqrt{n}(t - \theta)] \\ &= \tilde{\Phi}(\sqrt{n}(t - \theta)) \quad \text{where } \tilde{\Phi}(z) = \mathbb{P}_{Z \sim N(0,1)}[Z > z]. \end{aligned} \tag{3.1.2}$$

This test has size $\leq \alpha$ as long as under $\theta = 0$

$$\mathbb{P}_{\theta=0}[\phi(X) = 1] = \tilde{\Phi}(\sqrt{n}t) \leq \alpha \iff t \geq \frac{1}{\sqrt{n}} z^{(1-\alpha)}. \tag{3.1.3}$$

where $z^{(1-\alpha)}$ is the $(1-\alpha)$ - quantile of the $N(0,1)$ distribution.

To maximize power, choose $t = \frac{1}{\sqrt{n}} z^{(1-\alpha)}$.
Summary: the test is to reject H_0 when $\sqrt{n}\bar{X} > z^{(1-\alpha)}$.
The power is

$$\begin{aligned} \beta(\theta) &= \mathbb{P}_\theta[\phi(X) = 1] \\ &= \tilde{\Phi}(\sqrt{n}(t-\theta)) \\ &= \tilde{\Phi}(\sqrt{n}(\frac{1}{\sqrt{n}} z^{(1-\alpha)} - \theta)) \\ &= \tilde{\Phi}(z^{(1-\alpha)} - \sqrt{n}\theta) \\ &= 1 - \tilde{\Phi}(\sqrt{n}\theta - z^{(1-\alpha)}). \end{aligned} \tag{3.1.4}$$

Note: For $\theta \asymp \frac{1}{\sqrt{n}}$, $\beta(\theta)$ is of constant order. For any fixed $\theta > 0$, and $n \to \infty$, $\beta(\theta) \to 1$ exponentially fast.

Definition 3.1.3 (P-value)**.** Consider a non-randomized test ϕ_α for each $\alpha \in (0,1)$, where each ϕ_α has size $\leq \alpha$. Let $S_\alpha = \{x \in \mathcal{X} : \phi_\alpha(x) = 1\}$ be the rejection region. Suppose that

$$\alpha \leq \alpha' \implies S_\alpha \subseteq S_{\alpha'}. \tag{3.1.5}$$

Then the *p-value* of the observed data X is

$$\hat{p}(X) = \inf\{\alpha : X \in S_\alpha\}, \tag{3.1.6}$$

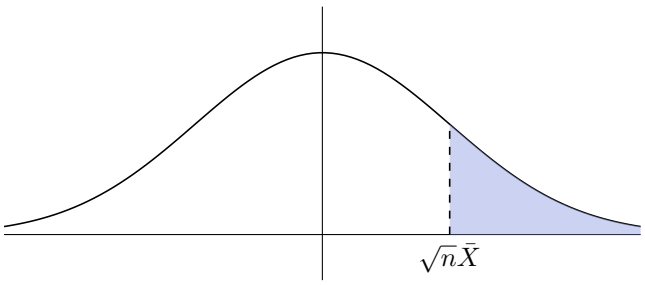
which is the smallest level α at which we reject H_0.

Example 3.1.2. Consider the test in Example 3.1.1

$$\phi(X) = \begin{cases} 1 & \sqrt{n}\bar{X} > z^{(1-\alpha)} \\ 0 & \sqrt{n}\bar{X} \leq z^{(1-\alpha)} \end{cases} \tag{3.1.7}$$

The p-value is

$$\begin{aligned} \hat{p} &= \inf\{\alpha : \sqrt{n} > z^{(1-\alpha)}\} \\ &= \text{ the value } \alpha \text{ where } \sqrt{n}\bar{X} = z^{(1-\alpha)} \\ &= \tilde{\Phi}(\sqrt{n}\bar{X}). \end{aligned} \tag{3.1.8}$$



Proposition 3.1.1. Suppose that p-value is well-defined:

$$\alpha \leq \alpha' \implies S_\alpha \subseteq S_{\alpha'} \tag{3.1.9}$$

and that each test is valid:

$$\sup_{\theta \in \Omega_0} \mathbb{P}_\theta[X \in S_\alpha] \leq \alpha \quad \text{for all } \alpha \in (0,1). \tag{3.1.10}$$

(i) Under every $\theta \in \Omega_0$,

$$\mathbb{P}_\theta[\hat{p} \leq u] \leq u \quad \text{for all } u \in [0,1]. \tag{3.1.11}$$

(ii) For any $\theta \in \Omega_0$, if $\mathbb{P}_\theta[X \in S_\alpha] = \alpha$ for every $\alpha \in (0,1)$, then

$$\hat{p} \sim Uniform([0,1]) \text{ under } \theta \quad (\text{i.e., } \mathbb{P}_\theta[\hat{p} \leq u] = u). \tag{3.1.12}$$

Proof.

(i) If $\hat{p} \leq u$, then $x \in S_v$ for all $v > u$. Thus

$$\mathbb{P}_\theta[\hat{p} \leq u] \leq \mathbb{P}_\theta[X \in S_v] \leq v. \tag{3.1.13}$$

Take $v \searrow u$ and we get the result.

(ii) If $X \in S_u$, then $\hat{p} \leq u$. Then

$$u = \mathbb{P}_\theta[X \in S_u] \leq \mathbb{P}_\theta[\hat{p} \leq u]. \tag{3.1.14}$$

Combining with (i), we must have $u = \mathbb{P}_\theta[\hat{p} \leq u]$.

□

Example 3.1.3. In Example 3.1.1 and Example 3.1.2, at $\theta = 0$, we did have $\mathbb{P}_{\theta=0}[\phi_\alpha(X) = 1] = \alpha$. And indeed,

$$\begin{aligned} \mathbb{P}_{\theta=0}[\hat{p} \leq u] &= \mathbb{P}_{\theta=0}[\tilde{\Phi}(\sqrt{n}\bar{X}) \leq u] \\ &= \mathbb{P}_{\theta=0}[\sqrt{n}\bar{X} \geq z^{(1-u)}] \\ &= u. \\ \text{So } \hat{p} &\sim Uniform([0,1]). \end{aligned} \tag{3.1.15}$$

Suppose we consider the same test for

$$H_0 : \theta = 0 \text{ vs. } H_1 : \theta > 0.$$

At every $\theta < 0$,

$$\mathbb{P}_\theta[\sqrt{n}\bar{X} \geq z^{(1-u)}] < u \text{ with strict inequality.} \tag{3.1.16}$$

Definition 3.1.4 (Simple and composite)**.** A null or alternative hypothesis is *simple* if it consists of only a single distribution (i.e., $\Omega_0 = \{\theta_0\}$). Otherwise, it is *composite*.

3.2 Neyman-Pearson Lemma

Theorem 3.2.1 (Neyman-Pearson Lemma)**.** Suppose the null hypothesis and the alternative hypothesis are both simple. $\Omega_0 = \{\theta_0\}, \Omega_1 = \{\theta_1\}$ and $P_{\theta_0}, P_{\theta_1}$ have densities p_0, p_1 with common measure μ. Then for any $\alpha \in (0,1)$:

(i) There exists a possibly randomized test $\phi : \mathcal{X} \to [0,1]$ such that

$$\mathbb{E}_{\theta_0}[\phi(X)] = \alpha \quad \text{test has size exactly } \alpha. \tag{3.2.1}$$

and

$$\phi(x) = \begin{cases} 1 & p_1(x) > k \cdot p_0(x) \\ 0 & p_1(x) < k \cdot p_0(x) \end{cases} \quad \text{for some } k \equiv k(\alpha). \tag{3.2.2}$$

(ii) Any test satisfying both condition of (i) is most powerful: For any other test ϕ' with size $\mathbb{E}_{\theta_0}[\phi(X)] \le \alpha$, we have

$$\begin{aligned} &\mathbb{E}_{\theta_1}[\phi'(X)] \le \mathbb{E}_{\theta_1}[\phi(X)], \\ &\text{power of } \phi' \le \text{power of } \phi. \end{aligned} \tag{3.2.3}$$

(iii) Uniqueness: Any test which is most powerful must satisfy (3.2.2) for some $k \equiv k(\alpha)$ $(\mu - a.e.)$.

Proof.

(i) For x where $p_0(x) \neq 0$, define $L(x) = \frac{p_1(x)}{p_0(x)} \in [0, +\infty)$. Let $F(k) = \mathbb{P}_{\theta_0}[L(X) \le k]$ be the CDF of $L(X)$ under H_0. $F(k)$ is non-decreasing, right-continuous. There exists k where

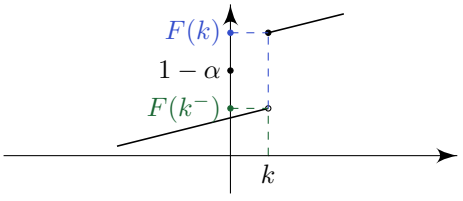
$$F(k) \ge 1 - \alpha \text{ and } F(k^-) \le 1 - \alpha.$$

Note that $\mathbb{P}_{\theta_0}[L(X) = k] = F(k) - F(k^-)$. Let

$$\phi(x) = \begin{cases} 1 & p_1(x) > k \cdot p_0(x) \\ \frac{F(k)-(1-\alpha)}{F(k)-F(k^-)} & p_1(x) = k \cdot p_0(x) \\ 0 & p_1(x) < k \cdot p_0(x) \end{cases} \tag{3.2.4}$$

This satisfies (3.2.2) and

$$\begin{aligned} \mathbb{E}_{\theta_0}[\phi(X)] &= P_{\theta_0}[L(X) > k] + \frac{F(k) - (1-\alpha)}{F(k) - F(k^-)} \cdot \mathbb{P}_{\theta_0}[L(X) = k] \\ &= P_{\theta_0}[L(X) > k] + \frac{F(k) - (1-\alpha)}{F(k) - F(k^-)} \cdot (F(k) - F(k^-)) \\ &= 1 - F(k) + F(k) - (1-\alpha) \\ &= \alpha. \end{aligned} \tag{3.2.5}$$



(ii) Let ϕ satisfy the conditions of (i). Consider

$$\begin{aligned}\mathcal{L}(\phi') &\equiv \mathbb{E}_{\theta_1}[\phi'(X)] - k \cdot \mathbb{E}_{\theta_0}[\phi'(X)] \\ &= \int \phi'(x)(p_1(x) - k \cdot p_0(x))\mathrm{d}\mu(x).\end{aligned} \tag{3.2.6}$$

For each $x \in \mathcal{X}$, $\phi'(x)(p_1(x) - k \cdot p_0(x))$ is maximized at

$$\phi'(x) = \begin{cases} 1 & p_1(x) - k \cdot p_0(x) > 0 \\ 0 & p_1(x) - k \cdot p_0(x) < 0. \end{cases} \implies \mathcal{L}(\phi') \le \mathcal{L}(\phi) \text{ for } \textit{any} \text{ test } \phi'.$$

By (i) ϕ has size exactly α, if ϕ' has size $\le \alpha$, then

$$\begin{aligned}&\mathbb{E}_{\theta_1}[\phi'(X)] - k\alpha \le \mathcal{L}(\phi') \le \mathcal{L}(\phi) = \mathbb{E}_{\theta_1}[\phi(X)] - k\alpha, \\ &\mathbb{E}_{\theta_1}[\phi'(X)] \le \mathbb{E}_{\theta_1}[\phi(X)].\end{aligned} \tag{3.2.7}$$

So ϕ is the most powerful test.

(iii) If ϕ *doesn't* satisfy (3.2.2), then

$$\mathcal{L}(\phi') < \mathcal{L}(\phi) \tag{3.2.8}$$

holds strictly. Then also

$$\mathbb{E}_{\theta_1}[\phi'(X)] < \mathbb{E}_{\theta_1}[\phi(X)] \tag{3.2.9}$$

holds strictly. So ϕ' is not the most powerful test.

□

Remark 3.2.1. Informally, Let $L(x) = \frac{p_1(x)}{p_0(x)} \in [0, +\infty)$ be the *likelihood ratio statistic*.

- There exists a most powerful test of H_0 vs. H_1 which rejects H_0 when $L(x)$ is large.
- If the distribution of $L(x)$ is continuous, then the probability that $L(x) = k$ is 0. The most powerful test is unique.
- If $L(x)$ is discrete, ϕ may need to randomize when $L(x) = k$ to achieve maximal power. The most powerful test is unique up to how it performs this randomization.

Example 3.2.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$. Test

$$H_0 : \theta = 0 \text{ vs. } H_1 : \theta = \theta_1 > 0$$

for some prespecified, known value θ_1. Then

$$\begin{aligned}p_0(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}, \\ p_{\theta_1}(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta_1)^2}{2}}.\end{aligned}$$

$$\begin{aligned} L(x) &= \frac{p_{\theta_1}(x)}{p_0(x)} \\ &= \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_1)^2}{2}+\frac{x_i^2}{2}} \\ &= \exp\left(\left(\sum_{i=1}^{n} x_i\right)\cdot\theta_1 - \frac{n\theta_1^2}{2}\right) \\ &= \exp\left(n\theta_1\cdot\bar{x} - \frac{n\theta_1^2}{2}\right). \end{aligned}$$

By Neyman-Pearson Lemma, the most powerful level-α test rejects H_0 for large $L(X)$ and we need to choose $k(\alpha)$ such that

$$\mathbb{P}_{\theta=0}[L(X) > k(\alpha)] = \alpha \iff \mathbb{P}_{\theta=0}[\bar{X} > \tilde{k}(\alpha)] = \alpha.$$
$$\text{where } \tilde{k}(\alpha) \equiv \frac{\log k(\alpha) + \frac{n\theta_1^2}{2}}{n\theta_1}.$$

Under H_0, $\bar{X} \sim N(0, \frac{1}{n})$. So we should pick $\tilde{k}(\alpha)$ to be $\frac{1}{\sqrt{n}}z^{(1-\alpha)}$ to guarantee $\mathbb{P}_{\theta=0}[\bar{X} > \tilde{k}(\alpha)] = \alpha$. So the most powerful test is

$$\phi(X) = \begin{cases} 1 & \sqrt{n}\bar{X} > z^{(1-\alpha)} \\ 0 & \sqrt{n}\bar{X} < z^{(1-\alpha)} \end{cases} \tag{3.2.10}$$

Remark 3.2.2. The form of $\phi(X)$ *does not* depend on θ, other than the fact that $\theta_1 > 0$. This test is *uniformly* most powerful against $H_1 : \theta > 0$ (i.e., the most powerful test is the same for all $\theta_1 > 0$).

Example 3.2.2. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. Test

$$H_0 : \theta = \theta_0 = \frac{1}{2} \text{ vs. } H_1 : \theta = \theta_1 > \frac{1}{2}$$

for some prespecified, known value θ_1. Then

$$\begin{aligned} p_{\theta_0}(x) &= \prod_{i=1}^{n}\frac{1}{2} = (\frac{1}{2})^n, \\ p_{\theta_1}(x) &= \prod_{i=1}^{n}\theta_1^{x_i}(1-\theta_1)^{1-x_i} = \theta_1^s(1-\theta_1)^{n-s} \quad \text{where } s = \sum_{i=1}^{n} x_i, \\ L(x) &= \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = 2^n\cdot\theta_1^s(1-\theta_1)^{n-s} = 2^n\cdot(1-\theta_1)^n\cdot\left(\frac{\theta_1}{1-\theta_1}\right)^s. \end{aligned}$$

By Neyman-Pearson Lemma, the most powerful level-α test rejects H_0 for large $L(X)$. When $\theta_1 > \frac{1}{2}$, $L(X)$ is increasing in S. So equivalently,

$$\phi(X) = \begin{cases} 1 & S > \tilde{k}(\alpha) \\ \tilde{c}(\alpha) & S = \tilde{k}(\alpha) \\ 0 & S < \tilde{k}(\alpha). \end{cases}$$

We are going to choose $\tilde{k}(\alpha), \tilde{c}(\alpha)$ such that

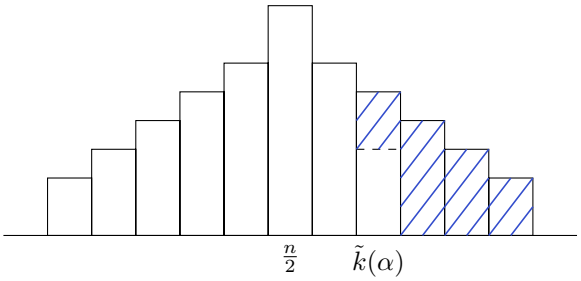
$$\mathbb{P}_{\theta=\theta_0=\frac{1}{2}}[S > \tilde{k}(\alpha)] + \tilde{c}(\alpha)\cdot\mathbb{P}_{\theta=\theta_0=\frac{1}{2}}[S = \tilde{k}(\alpha)] = \alpha.$$

Under H_0, $S \sim Binomial(n, \frac{1}{2})$. Let $\tilde{k}(\alpha)$ be such that

$$\mathbb{P}[Binomial(n, \frac{1}{2}) \geq \tilde{k}(\alpha)] \geq \alpha,$$
$$\mathbb{P}[Binomial(n, \frac{1}{2}) \geq \tilde{k}(\alpha) + 1] \leq \alpha.$$

Let $G(k) = \mathbb{P}[S \geq k]$ for $S \sim Bionomial(n, \frac{1}{2})$. Therefore

$$\phi(X) = \begin{cases} 1 & S > \tilde{k}(\alpha) \\ \frac{\alpha - G(\tilde{k}(\alpha)+1)}{G(\tilde{k}(\alpha)) - G(\tilde{k}(\alpha)+1)} & S = \tilde{k}(\alpha) \\ 0 & S < \tilde{k}(\alpha). \end{cases} \tag{3.2.11}$$



Remark 3.2.3. Again, this form doesn't depend on the exact value of θ_1. So $\phi(X)$ is the universal most powerful test against $\theta > \frac{1}{2}$.

Theorem 3.2.2. Let $p_\eta(x) = \exp(\eta \cdot T(x) - A(\eta)) \cdot h(x)$. For any fixed $\eta = \eta_0$, there is a uniformly most powerful (UMP) test of

$$H_0 : \eta = \eta_0 \text{ vs. } H_1 : \eta_1 > \eta_0$$

given by

$$\phi(X) = \begin{cases} 1 & T(X) > k \\ c & T(X) = k \\ 0 & T(X) < k. \end{cases}$$

where k, c are chosen such that $\mathbb{P}_{\eta_0}[T(X) > k] + c \cdot \mathbb{P}_{\eta_0}[T(X) = k] = \alpha$. (3.2.12)

Proof. Consider any *fixed* alternative $\eta_1 > \eta_0$.

$$L(x) = \frac{p_{\eta_1}(x)}{p_{\eta_0}(x)} = \exp((\eta_1 - \eta_0) \cdot T(x) - A(\eta_1) + A(\eta_0)).$$

Since $\eta_1 > \eta_0$, $L(x)$ is increasing in $T(x)$. The result follows from Neyman-Pearson Lemma. □

Remark 3.2.4. If we want to test

$$H_0 : \eta = \eta_0 \text{ vs. } H_1 : \eta_1 < \eta_0,$$

the most powerful test would reject H_0 for small $T(X)$. But for

$$H_0 : \eta = \eta_0 \text{ vs. } H_1 : \eta_1 \neq \eta_0,$$

there is *no* UMP test. So we are going to "balance" power against different alternatives of interest.

Several themes:

- Consider a prior Λ on the alternative space Ω_1, and maximize the average power $\int_{\Omega_1} \beta(\theta) \mathrm{d}\Lambda(\theta)$.
- Maxmin: maximize the minimum power $\inf_{\theta \in \Omega_1} \beta(\theta)$.
- Sparse vs. dense alternatives.

3.3 Normal Means Testing

Setup: Suppose $X = (X_1, X_2, \cdots, X_d) \sim N(\theta, I) \in \mathbb{R}^d,\ \theta = (\theta_1, \theta_2, \cdots, \theta_d)$. Test

$$H_0 : \theta = 0 \text{ vs. } H_1 : \theta \neq 0.$$

Example 3.3.1. Consider a subset of sparse alternatives. For some known $\mu > 0$, we want to test

$$\begin{aligned} H_1' &: \{\theta : \theta_i = \mu \text{ for one coordinate } i, \text{ all other coordinates are } 0\} \\ &= \{(\mu, 0, 0, \cdots, 0), (0, \mu, 0, \cdots, 0), \cdots, (0, \cdots, 0, 0, \mu)\} \\ &= \{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(n)}\}. \end{aligned}$$

Question 3.3.1. How to maximize power against $\theta^{(1)}$?

Answer 3.3.1. Apply Neyman-Pearson Lemma,

$$\begin{aligned} p_0(x) &= \left(\frac{1}{\sqrt{2\pi}}\right)^d \cdot e^{-\frac{x_1^2}{2} - \frac{x_2^2}{2} - \cdots - \frac{x_d^2}{2}}, \\ p_{\theta^{(1)}}(x) &= \left(\frac{1}{\sqrt{2\pi}}\right)^d \cdot e^{-\frac{(x_1-\mu)^2}{2} - \frac{x_2^2}{2} - \cdots - \frac{x_d^2}{2}}. \\ L(x) &= \frac{p_{\theta^{(1)}}(x)}{p_0(x)} = e^{-\frac{(x_1-\mu)^2}{2} + \frac{x_1^2}{2}} = e^{\mu x_1 - \frac{\mu^2}{2}}. \end{aligned}$$

For $\mu > 0$, $L(X)$ is increasing in S. So the most powerful test rejects H_0 for large x_1.

Question 3.3.2. Take Λ as the uniform prior on $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(d)}\}$. How to maximize the average power against $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(d)}$ under Λ?

Answer 3.3.2.
Idea: the average power against Λ is *equal* to the power against the alternative

$$X \sim p_1(x) = \int_{\Omega_1} p_\theta(x) \mathrm{d}\Lambda(\theta), \tag{3.3.1}$$

the marginal distribution of X under Λ. So we can derive the test maximizing this average power using the Neyman-Pearson Lemma.

Proof.

$$\begin{aligned}\int_{\Omega_1}\beta(\theta)\mathrm{d}\Lambda(\theta) &= \int_{\Omega_1}\mathbb{E}_\theta[\phi(X)]\mathrm{d}\Lambda(\theta)\\ &= \int_{\Omega_1}\int_{\mathcal{X}}\phi(x)\cdot p_\theta(x)\mathrm{d}x\mathrm{d}\Lambda(\theta)\\ &= \int_{\mathcal{X}}\phi(x)\cdot\left(\int_{\Omega_1}p_\theta(x)\mathrm{d}\Lambda(\theta)\right)\mathrm{d}x \equiv \int_{\mathcal{X}}\phi(x)\cdot p_1(x)\mathrm{d}x.\end{aligned} \tag{3.3.2}$$

□

In our example,

$$\begin{aligned}p_0(x) &= \left(\frac{1}{\sqrt{2\pi}}\right)^d\cdot e^{-\frac{x_1^2}{2}-\frac{x_2^2}{2}-\cdots-\frac{x_d^2}{2}},\\ p_1(x) &= \frac{1}{d}(p_{\theta^{(1)}}+p_{\theta^{(2)}}+\cdots+p_{\theta^{(d)}})\\ &= \frac{1}{d}\left(\left(\frac{1}{\sqrt{2\pi}}\right)^d\cdot e^{-\frac{(x_1-\mu)^2}{2}-\frac{x_2^2}{2}-\cdots-\frac{x_d^2}{2}}+\cdots+\left(\frac{1}{\sqrt{2\pi}}\right)^d\cdot e^{-\frac{x_1^2}{2}-\frac{x_2^2}{2}-\cdots-\frac{(x_d-\mu)^2}{2}}\right).\\ L(x) &= \frac{p_1(x)}{p_0(x)}\\ &= \frac{1}{d}\cdot(e^{\mu x_1-\frac{\mu^2}{2}}+e^{\mu x_2-\frac{\mu^2}{2}}+\cdots+e^{\mu x_d-\frac{\mu^2}{2}})\\ &= \frac{1}{d}\cdot e^{-\frac{\mu^2}{2}}\cdot\sum_{i=1}^d e^{\mu x_i}\equiv\frac{1}{d}\cdot e^{-\frac{\mu^2}{2}}\cdot T(x)\end{aligned} \tag{3.3.3}$$

By Neyman-Pearson Lemma, the most powerful test of p_0 vs. p_1 rejects H_0 for large value of $L(X)$. Since $L(X)$ is increasing in $T(X)$, the most powerful test rejects for large $T(X)$.

Question 3.3.3. How to maximize the minimum power over $\{\theta^{(1)},\theta^{(2)},\cdots,\theta^{(d)}\}$.

Answer 3.3.3. Consider the

Proposition 3.3.1. Consider any prior Λ on Ω_1. If the test is maximizing the average power under Λ has constant power

$$\beta(\theta)=\int_{\Omega_1}\beta(\theta)\mathrm{d}\Lambda(\theta)\text{ for all }\theta\in\Omega_1, \tag{3.3.4}$$

then the test is also maximin.

Proof. Call this test ϕ. By assumption, for any test ϕ',

$$\begin{aligned}\inf_{\theta\in\Omega_1}\beta_{\phi'}(\theta) &\le \int_{\Omega_1}\beta_{\phi'}(\theta)\mathrm{d}\Lambda(\theta)\\ &\le \int_{\Omega_1}\beta_{\phi}(\theta)\mathrm{d}\Lambda(\theta)\\ &= \inf_{\theta\in\Omega_1}\beta_{\phi}(\theta).\end{aligned} \tag{3.3.5}$$

□

By Example 3.3.2 and (3.3.3), the test maximizing the average power under uniform prior Λ over $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(d)}\}$ was to reject H_0 for large $\sum_{i=1}^{d} e^{\mu X_i}$. By symmetry, this test has equal power against $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(d)}$. So this test is also maximin.

Question 3.3.4. How to set a rejection threshold for this test?

Answer 3.3.4. We want to set as $(1-\alpha)$-quantile of $\sum_{i=1}^{d} e^{\mu X_i}$ under $H_0 : \theta = 0$. If μ is large:

$$\sum_{i=1}^{d} e^{\mu X_i} \approx e^{\mu \cdot \max_{i=1}^{d} X_i} \equiv e^{\mu \cdot M(X)}.$$

This is increasing in $M(X)$. So we want to choose $k(\alpha)$ such that

$$\mathbb{P}_{\theta=0}[M(X) > k(\alpha)] \le \alpha.$$

We can use a conservative union bond

$$\begin{aligned}
\mathbb{P}_{\theta=0}[M(X) > k(\alpha)] &= \mathbb{P}_{\theta=0}\left[\bigcup_{i=1}^{d} \{X_i > k(\alpha)\}\right] \\
&\le \sum_{i=1}^{d} \mathbb{P}_{\theta=0}[X_i > k(\alpha)] \\
&= d \cdot \tilde{\Phi}(k(\alpha)).
\end{aligned}$$

To ensure it's $\le \alpha$, set $k(\alpha)$ to be $z^{(1-\frac{\alpha}{d})}$, the $(1-\frac{\alpha}{d})$-quantile of $N(0,1)$. Then

$$d \cdot \tilde{\Phi}(k(\alpha)) = d \cdot \frac{\alpha}{d} = \alpha.$$

This means we test $\theta_i = 0$ individually using X_i at the significance level $\frac{\alpha}{d}$. We reject H_0 if at least one such test rejects $\theta_i = 0$. This is the *Bonferroni procedure*.

Example 3.3.2. Consider a subset of alternatives. For some given $c > 0$, we want to test

$$H_1'' : \{\theta : \|\theta\| = c\}.$$

Question 3.3.5. Consider the uniform prior Λ on $\Omega_1 = \{\theta : \|\theta\| = c\}$. What is the test maximizing the average power?

Answer 3.3.5. The marginal distribution of X under Λ is

$$\begin{aligned}
p_1(x) &= \int_{\Omega_1} p_\theta(x) \mathrm{d}\Lambda(\theta) \\
&= \int_{\Omega_1} \left(\frac{1}{\sqrt{2\pi}}\right)^d \cdot e^{-\frac{\sum_{i=1}^{d} (x_i - \theta_i)^2}{2}} \mathrm{d}\Lambda(\theta) \\
&= \int_{\Omega_1} \left(\frac{1}{\sqrt{2\pi}}\right)^d \cdot e^{-\frac{\sum_{i=1}^{d} x_i^2}{2} + \sum_{i=1}^{d} x_i\theta_i - \frac{\sum_{i=1}^{d} \theta_i^2}{2}} \mathrm{d}\Lambda(\theta) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^d \cdot e^{-\frac{\sum_{i=1}^{d} x_i^2}{2} - \frac{c^2}{2}} \cdot \int_{\Omega_1} e^{x^T\theta} \mathrm{d}\Lambda(\theta).
\end{aligned}$$

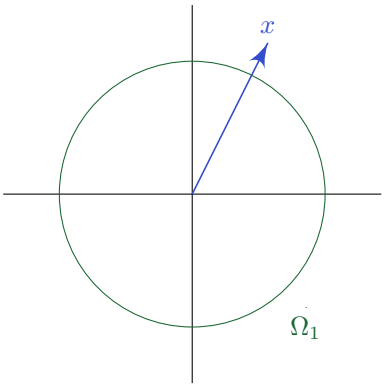
Recall that

$$p_0(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \cdot e^{-\frac{\sum_{i=1}^d x_i^2}{2}}.$$

So

$$L(x) = e^{-\frac{c^2}{2}} \cdot \int_{\Omega_1} e^{x^T\theta} \mathrm{d}\Lambda(\theta),$$

where $\int_{\Omega_1} e^{x^T\theta}\mathrm{d}\Lambda(\theta)$ is invariant to rotations of x.



So we pick $\tilde{x} = (\|x\|, 0, \cdots, 0)$.

$$\begin{aligned} L(x) = L(\tilde{x}) &= e^{-\frac{c^2}{2}} \cdot \int_{\Omega_1} e^{\theta_1 \cdot \|x\|}\mathrm{d}\Lambda(\theta) \\ &= e^{-\frac{c^2}{2}} \cdot \int_{-c}^{c} e^{\theta_1 \cdot \|x\|}\pi(\theta_1)\mathrm{d}\theta_1 \end{aligned}$$

where $\pi(\theta_1)$ is the density of θ_1 when $\theta \sim \Lambda$.

Observe that $\pi(\theta_1) = \pi(-\theta_1)$. So

$$L(x) = e^{-\frac{c^2}{2}} \cdot \int_0^c \left(e^{\theta_1 \cdot \|x\|} + e^{-\theta_1 \cdot \|x\|}\right)\pi(\theta_1)\mathrm{d}\theta_1. \tag{3.3.6}$$

Since $y \to e^{\theta_1 y} + e^{-\theta_1 y}$ is increasing over $y > 0$ for any $\theta > 0$, $L(X)$ is monotonically increasing in $\|X\|$. So the Neyman-Pearson Lemma is equivalent to rejecting H_0 for large $\|X\|^2$.

Under H_0

$$\|X\|^2 \sim \mathcal{X}_d^2.$$

Let $k(\alpha)$ be the $(1-\alpha)$-quantile of $\mathcal{X}_d^2$. Then the test is

$$\phi(X) = \begin{cases} 1 & \|X\|^2 > k(\alpha) \\ 0 & \|X\|^2 < k(\alpha). \end{cases} \tag{3.3.7}$$

This is the $\mathcal{X}^2$-test of H_0.

3.4 Multiple Hypotheses Testing

Setup: n total hypotheses, e.g., $\theta = (\theta_1, \theta_2, \cdots, \theta_n) \in \mathbb{R}^n$.

$$H_{0,i} : \theta_i = 0 \quad \text{(gene i is null)}$$
$$H_{1,i} : \theta_i > 0 \quad \text{(gene i is active)}.$$

We reduce data down to the n p-values $p_1, p_2, \cdots, p_n$.

Recall: In Proposition 3.1.1, under the i^{th} *null* hypothesis,

$$\mathbb{P}[p_i \leq u] \leq u \quad \forall u \in [0, 1]. \tag{3.4.1}$$

Types of outcomes:

	H_0 accepted	H_0 rejected	Total
H_0 true	U	V	n_0
H_0 false	T	S	$n - n_0$
Total	$n - R$	R	n.

We treat n_0, n as fixed (i.e., which nulls are true and which nulls are false is determined). U, V, T, S, R are all random. U, V, T, S are unobserved.
Previously, we were controlling Type I error under the "global null hypothesis".

$$H_0 : \text{ All } H_{0,i} \text{ are true for } i = 1, 2, \cdots$$

In practice, we'd like to also have error controlled in setting where some $H_{0,i}$ are true and some are false.

Definition 3.4.1 (Family-wise error rate)**.** *Family-wise error rate* (FWER) is the probability that we falsely reject at least one true null.

$$\text{FWER} = \mathbb{P}[V \geq 1]. \tag{3.4.2}$$

Definition 3.4.2 (False-discovery rate)**.** *False-discovery rate* (FDR) is the expected proportion of rejections that are false.

$$\text{FDR} = \mathbb{E}\left[\frac{V}{\max(R, 1)}\right]. \tag{3.4.3}$$

A procedure controls FWER/FDR at level α if FWER/FDR $\leq \alpha$ for *any* configuration of true and false null hypotheses.

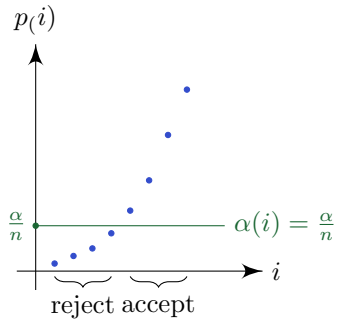
Question 3.4.1. How to control FWER at level α?

Answer 3.4.1. The *Bonferroni procedure* and the *Holm's procedure*.

Definition 3.4.3 (Bonferroni procedure)**.** The *Bonferroni procedure* rejects $H_{0,i}$ if

$$p_i \leq \frac{\alpha}{n}. \tag{3.4.4}$$

Note 3.4.1. If we order the p-values $p_{(n)} \leq p_{(n)} \leq \cdots \leq p_{(n)}$, the Bonferroni procedure compares each $p_{(i)}$ to the constant threshold $\frac{\alpha}{n}$.



Proposition 3.4.1. The Bonferroni procedure controls FWER at level α.

Proof. Let $i = 1, 2, \cdots, n_0$ correspond to the true nulls, $i = n_0 + 1, n_0 + 2, \cdots, n$ correspond to the false nulls.

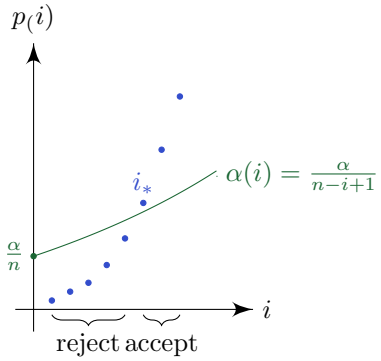
$$\begin{aligned} \text{FWER} &= \mathbb{P}[V \geq 1] \\ &= \mathbb{P}\left[\sum_{i=1}^{n_0} \{p_i \leq \frac{\alpha}{n}\}\right] \\ &\leq \sum_{i=1}^{n_0} \mathbb{P}\left[p_i \leq \frac{\alpha}{n}\right] \\ &\leq \sum_{i=1}^{n_0} \frac{\alpha}{n} = \frac{n_0 \alpha}{n} \leq \alpha. \end{aligned} \tag{3.4.5}$$

□

Definition 3.4.4 (Holm's procedure)**.** Let

$$\begin{aligned} \alpha(i) &= \frac{\alpha}{n - i + 1} \\ i_* &= \min\{i : p_{(i)} > \alpha(i)\}. \end{aligned} \tag{3.4.6}$$

The *Holm's procedure* rejects the hypotheses corresponding to $p_{(1)}, p_{(2)}, \cdots, p_{(i_*-1)}$.



Proposition 3.4.2. The Holm's procedure controls FWER at level α.

Proof. Let the ordered p-values be $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$. Let i_0 be such that $p_{(i_0)}$ is the first true null, so $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(i_0-1)}$ correspond to false nulls. If $V \geq 1$ (i.e., the Holm's procedure makes a false rejection), then it must reject the hypothesis for $p_{(i_0)}$

$$p_{(i_0)} \leq \alpha(i_0).$$

Note that $i_0 \leq n - n_0 + 1$, so

$$\alpha(i_0) = \frac{\alpha}{n - i_0 + 1} \leq \frac{\alpha}{n_0}.$$

Thus

$$\begin{aligned} \text{FWER} = \mathbb{P}[V \geq 1] &\leq \mathbb{P}[p_{(i_0)} \leq \frac{\alpha}{n_0}] \\ &= \mathbb{P}\left[\bigcup_{i \in \{\text{true nulls}\}} \{p_i \leq \frac{\alpha}{n_0}\}\right] \\ &\leq \sum_{i \in \{\text{true nulls}\}} \mathbb{P}\left[p_i \leq \frac{\alpha}{n_0}\right] \\ &\leq \sum_{i=1}^{n_0} \frac{\alpha}{n_0} =\leq \frac{n_0 \alpha}{n_0} = \alpha. \end{aligned} \tag{3.4.7}$$

□

Remark 3.4.1. the Holm's procedure compares $p_{(1)}$ to $\alpha(1) = \frac{\alpha}{n}$, the Bonferroni level. However, if $p_{(1)} < \frac{\alpha}{n}$, then it compares $p_{(2)}$ to $\alpha(1) = \frac{\alpha}{n-1}$, a slightly less conservative threshold. Each subsequent $p_{(i)}$ is compared to a less conservative threshold than the one for $p_{(i-1)}$. So the Holm's procedure is strictly more powerful than the Bonferroni procedure as the rejected hypotheses must contain those rejected by the Bonferroni procedure.

Remark 3.4.2. If a procedure controls FWER at level α, it also controls FDR at level α. Since $V \leq R$

$$\frac{V}{\max(R,1)} = \begin{cases} 0 & V = R = 0 \\ \frac{V}{R} & V > 0 \end{cases} \leq \mathbb{1}\{V \geq 1\}.$$

Then

$$\begin{aligned} \text{FDR} = \mathbb{E}\left[\frac{V}{\max(R,1)}\right] &\leq \mathbb{E}[\mathbb{1}\{V \geq 1\}] \\ &= \mathbb{P}[V \geq 1] = \text{FWER}. \end{aligned} \tag{3.4.8}$$

However, in general FDR is a looser crterion. Suppose there are 100000 hypotheses, 1000 of which are false nulls. We can reject all of these correctly, in addition to false rejecting ≈ 50 true nulls and still control FDR at level $\alpha = 0.05$.

Question 3.4.2. How to control FDR at level α?

Answer 3.4.2. The *Benjamini-Hochberg* procedure.

Idea: Suppose we reject all p-values $\leq t$. Null p-values should be (at worst) uniformly distributed over $(0,1)$. So if there are n_0 true nulls, we expect

$$\begin{aligned} V &\approx n_0 t \leq nt \\ \frac{V}{R} &\leq \frac{nt}{\text{number of } \{p\text{-values} \leq t\}} \equiv \tilde{\text{FDR}}(t) \end{aligned}$$

$\tilde{\text{FDR}}(t)$ gives us a (possibly conservative) idea of the FDR of this procedure.

To maximize power subject to FDR $\leq \alpha$, consider rejecting p-values $\leq t_*$ for

$$\begin{aligned} t_* &= \max\{t \in (0,1) : \tilde{\text{FDR}}(t)\} \\ &= \max\{t : t \leq \frac{\alpha}{n} \cdot \text{number of } \{p\text{-values} \leq t\}\}. \end{aligned}$$

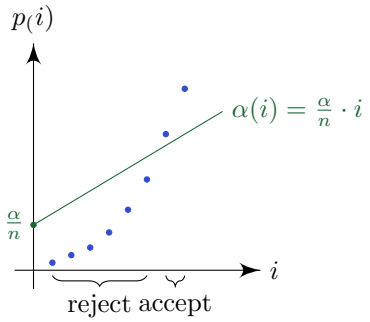
If this rejects $p_{(1)}, p_{(2)}, \cdots, p_{(i_*)}$ and accepts $p_{(i_*+1)}, p_{(i_*+2)}, \cdots, p_{(n)}$, then

$$\begin{aligned} &p_{(i_*)} \leq t_* < p_{(i_*+1)} \\ &i_* = \max\{i : p_{(i)} \leq \frac{\alpha}{n} \cdot i\}. \end{aligned} \tag{3.4.9}$$

Definition 3.4.5. Let

$$\begin{aligned} \alpha(i) &= \frac{\alpha}{n} \cdot i \\ i_* &= \begin{cases} 0 & p_{(i)} > \alpha(i) \quad \forall i. \\ \max\{i : p_{(i)} \leq \alpha(i)\} & \text{otherwise.} \end{cases} \end{aligned} \tag{3.4.10}$$

The *Benjamini-Hochberg* procedure rejects the hypotheses for $p_{(1)}, p_{(2)}, \cdots, p_{(i_*)}$.



Proposition 3.4.3. Suppose the p-values $p_1, p_2, \cdots, p_n$ are independent. Then the Benjamini-Hochberg procedure controls FDR at level α.

Proof. Let $i = 1, 2, \cdots, n_0$ be the true nulls, $i = n_0 + 1, n_0 + 2, \cdots, n$ be the false nulls. Then

$$\begin{aligned} FDR &= \mathbb{E}\left[\frac{V}{\max(R,1)}\right] \\ &= \sum_{r=1}^{n} \mathbb{E}\left[\frac{V}{r} \cdot \mathbb{1}\{R = r\}\right] \\ &= \sum_{r=1}^{n} \frac{1}{r} \sum_{i=1}^{n_0} \mathbb{E}\left[\mathbb{1}\{\text{reject} H_{0,i}, R = r\}\right]. \end{aligned}$$

By definition of the Benjamini-Hochberg procedure (Definition 3.4.5),

$$R = r \iff p_{(r)} \le \frac{\alpha}{n} \cdot r \text{ and } p_{(s)} > \frac{\alpha}{n} \cdot s \quad \forall s > r.$$

Fix i and let $p^{(i)}_{(1)} \le p^{(i)}_{(2)} \le \cdots \le p^{(i)}_{(n-1)}$ be the ordering of the $n-1$ p-values not including p_i. Then

$$\text{reject } H_{0,i} \text{ and } R = r \iff p_i \le \frac{\alpha}{n} \cdot r,\ p^{(i)}_{(r-1)} \le \frac{\alpha}{n} \cdot r \text{ and } p^{(i)}_{(s)} > \frac{\alpha}{n} \cdot (s+1) \quad \forall s > r.$$

Let

$$\varepsilon^{(i)}_r = \left\{ p^{(i)}_{(r-1)} \le \frac{\alpha}{n} \cdot r,\ p^{(i)}_{(s)} > \frac{\alpha}{n} \cdot (s+1) \quad \forall s > r \right\} \text{ for } r = 1, 2, \cdots, n.$$

Observe that

- $\varepsilon^{(i)}_r$ is independent of p_i by the independence of p-values.
- The events $\varepsilon^{(i)}_1, \varepsilon^{(i)}_2, \cdots, \varepsilon^{(i)}_n$ are mutually exclusive for any fixed i. Let

$$j^{(i)}_* = \begin{cases} 0 & p^{(i)}_{(j)} > \frac{\alpha}{n} \cdot (j+1) \\ \max\{j : p^{(i)}_{(j)} \le \frac{\alpha}{n} \cdot (j+1)\} & \text{otherwise.} \end{cases}$$

  We have $\varepsilon^{(i)}_r = \{j^{(i)}_* = r - 1\}$.

So

$$\begin{aligned} FDR &= \sum_{i=1}^{n} \frac{1}{r} \sum_{i=1}^{n_0} \mathbb{P}\left[p_i \le \frac{\alpha}{n}\right] \cdot r\mathbb{P}[\varepsilon^{(i)}_r] \\ &\le \sum_{i=1}^{n} \frac{1}{r} \sum_{i=1}^{n_0} \frac{\alpha}{n} \cdot r\mathbb{P}[\varepsilon^{(i)}_r] \qquad (3.4.11) \\ &= \frac{\alpha}{n} \sum_{i=1}^{n_0} \sum_{i=1}^{n} \mathbb{P}[\varepsilon^{(i)}_r] = \frac{n_0}{n} \cdot \alpha \le \alpha. \end{aligned}$$

□

Note 3.4.2. In fact, if $p_i \sim Uniform(0,1)$ for each true null hypothesis $H_{0,i}$, then FDR $= \frac{n_0}{n}\alpha$ exactly.

Remark 3.4.3.

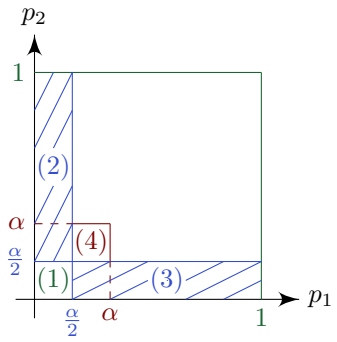
(i) $p_{(1)}$ is still compared to $\frac{\alpha}{n}$, the Bonferroni cutoff. Similar to the Holm's procedure in Remark 3.4.1, every subsequent comparison is less stringent.

(ii) This is less conservative than the Holm's procedure in two ways

- We have $\frac{\alpha}{n} \cdot i \ge \frac{\alpha}{n-i+1}$. In fact, for small i, $\frac{\alpha}{n-i+1} \approx \frac{\alpha}{n}$. So the Holm's procedure is not that much of an improvement over the Bonferroni procedure, but $\frac{\alpha}{n} \cdot i$ is larger by a factor of i.
- This looks at the *last* point where $p_{(i)} \le \alpha(i)$, whereas the Holm's procedure looks at the *first* point where $p_{(i)} > \alpha(i)$. In particular, the Benjamini-Hochberg procedure can reject some hypotheses even if $p_{(1)} > \frac{\alpha}{n}$ (the Bonferroni cutoff), as long as some later $p_{(i)}$ is less than $\frac{\alpha}{n} \cdot i$.

Question 3.4.3. What happens if p-values are *not* independent?

Example 3.4.1. Let $n = 2$ with both null hypotheses true. Let $p_1, p_2 \sim Uniform(0,1)$ but not necessarily independent.



Since both nulls are true.

$$\frac{V}{\max(R,1)} = \begin{cases} 0 & R = 0 \\ 1 & R = 1 \text{ or } 2. \end{cases}$$

Then

$$\text{FDR} = \mathbb{P}[R \geq 1] = \mathbb{P}[(1)] + \mathbb{P}[(2)] + \mathbb{P}[(3)] + \mathbb{P}[(4)].$$
$$\alpha = \mathbb{P}\left[p_1 \leq \frac{\alpha}{2}\right] + \mathbb{P}\left[p_2 \leq \frac{\alpha}{2}\right] = \mathbb{P}[(1)] + \mathbb{P}[(3)] + \mathbb{P}[(1)] + \mathbb{P}[(2)]$$

As

$$\mathbb{P}[(4)] \leq \mathbb{P}[\frac{p}{2} \leq p_1 \leq \alpha] = \frac{\alpha}{2}.$$

Then

$$FDR = \alpha + \mathbb{P}[(4)] - \mathbb{P}[(1)] \leq \frac{3\alpha}{2}. \tag{3.4.12}$$

We can have equality $FDR = \frac{3\alpha}{2}$ if we consider the density of (p_1, p_2) given by

$\frac{1}{1-\alpha}$	0	$\frac{1}{1-\alpha}(1 - \frac{\alpha}{2(1-\alpha)})$
0	$\frac{2}{\alpha}$	0
0	0	$\frac{1}{1-\alpha}$

We can check that p_1, p_2 are marginally $Uniform(0,1)$.

Theorem 3.4.1 (Benjamini-Yekutieli)**.** Let $S_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log n + 0.577$. The Benjamini-Hochberg procedure controls FDR at level $\alpha \cdot S_n$, under arbitrary dependence of $p_1, p_2, \cdots, p_n$.

Definition 3.4.6. $D \subseteq \mathbb{R}^n$ is increasing if $x \in D \implies y \in D$ for all $y \geq x$ (entrywise). $P_1, p_2, \cdots, p_n$ are PRDS on I_0 if for any increasing D and any $i \in I_0$, $x \to \mathbb{P}[(p_1, p_2, \cdots, p_n) \in D | p_i = p]$ is monotonically increasing in p.

Theorem 3.4.2 (Benjamini-Yekutieli)**.** If the p-values $p_1, p_2, \cdots, p_n$ are PRDS on I_0, the set of true nulls, then the Benjamini-Hochberg procedure controls FDR at level α.

Unit 2 Asymptotic Theory

Setting: Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta, \quad \theta \in \Omega \subseteq \mathbb{R}^k$. A k-parameter model for small fixed k. We want to study the behavior of estimators, tests as $n \to \infty$.

Reasons:

- Simplicity: Many estimators, tests which are difficult to study exactly for finite n become easy to study in the $n \to \infty$ limit.
- Universality: The phenomena we'll study hold more generally and become less dependent in the details of the specific distribution of the data. Important to understanding robustness against model misspecification.
- Accuracy: Oftentimes $n \gg k$ in practice, and the behavior as $n \to \infty$ is a good approximation of what occurs.

4 Convergence of Random Variables

4.1 Basic Convergence Theories

Theorem 4.1.1 (Central Limit Theorem)**.** If $X_1, X_2, \cdots, X_n \in \mathbb{R}$ are i.i.d. with mean μ, variance $\sigma^2 < \infty$, then

$$\sqrt{n} \cdot (\bar{X} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{\text{d}} N(0, \sigma^2). \tag{4.1.1}$$

Definition 4.1.1 (Converge in distribution)**.** Random variables Y_n *converge in distribution* or *converge in law* to Y if for any continuous, bounded function $f : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[f(Y_n)] \to \mathbb{E}[f(Y)]$ as $n \to \infty$. We'll write $Y_n \xrightarrow{\text{d}} Y$.

Equivalently: $Y_n \xrightarrow{\text{d}} Y$ if and only if the CDFs F_n of Y_n and F of Y satisfy $F_n(t) \to F(t)$ for every $t \in \mathbb{R}$ where F is continuous (i.e., $\mathbb{P}[Y = t] = 0$).

Theorem 4.1.2 (Multivariate Central Limit Theorem)**.** If $X_1, X_2, \cdots, X_n \in \mathbb{R}^k$ are i.i.d. with mean $\mu \in \mathbb{R}^k$, covariance $\Sigma \in \mathbb{R}^{k \times k}$, then

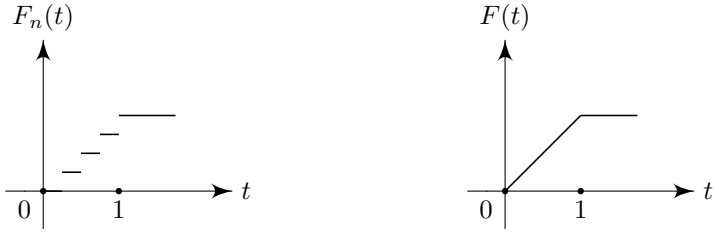
$$\sqrt{n} \cdot (\bar{X} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{\text{d}} N(0, \Sigma). \tag{4.1.2}$$

Definition 4.1.2 (Multivariate converge in distribution)**.** Random variables $Y_n \in \mathbb{R}^k$ *converge in distribution* or *converge in law* to $Y \in \mathbb{R}^k$ if for any continuous, bounded function $f : \mathbb{R}^k \to \mathbb{R}$, we have $\mathbb{E}[f(Y_n)] \to \mathbb{E}[f(Y)]$ as $n \to \infty$. We'll write $Y_n \xrightarrow{\text{d}} Y$.

Equivalently: $Y_n \xrightarrow{\text{d}} Y$ if and only $\mathbb{P}[Y_n \in E] \to \mathbb{P}[Y \in E]$ for every set $E \subset \mathbb{R}^k$ whose boundary ∂E satisfies $\mathbb{P}[Y \in \partial E] = 0$.

Note 4.1.1. This is a property of the *distributions* of Y_n and Y, not their real values. We'll also write $Y_n \xrightarrow{\text{d}} N(0, 1)$ or $F_n \xrightarrow{\text{d}} F$.

Example 4.1.1. Let Y_n have a discrete uniform distribution on $\{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n}{n}\}$, $Y \sim Uniform(0, 1)$. Then $Y_n \xrightarrow{\text{d}} Y$ as $n \to \infty$. (Check: $F_n(t) \to F(t)$ for all t).



But *not* true that $\mathbb{P}[Y_n \in E] \to \mathbb{P}[Y \in E]$ for every set E. E.g., take $E = \{\text{rational numbers}\}$, $\mathbb{P}[Y_n \in E] = 1 \ \forall n$ but $\mathbb{P}[Y \in E] = 0$. ($\partial E = [0, 1]$, $\mathbb{P}[Y \in \partial E] = 1$).

Proposition 4.1.1 (Convergence of Quantiles)**.** Let $\alpha \in (0, 1)$ and let $F^{(\alpha)}$ be the α^{th} quantile of F. If F is continuous and strictly increasing at $F^{(\alpha)}$.

$$F_n \xrightarrow{\text{d}} F \implies F_n^{(\alpha)} \to F^{(\alpha)}. \tag{4.1.3}$$

Proof.

- Take an increasing sequence $t_i \nearrow F^{(\alpha)}$ where F is continuous at each t_i. We have
$$\lim_{n\to\infty} F_n(t_i) = F(t_i) < \alpha,$$
so for sufficiently large n, $F_n(t_i) < \alpha$, implying $F_n^{(\alpha)} \geq t_i$.
- Take a decreasing sequence $s_i \searrow F^{(\alpha)}$ where F is continuous at each s_i. We have
$$\lim_{n\to\infty} F_n(s_i) = F(s_i) > \alpha,$$
so for sufficiently large n, $F_n(s_i) > \alpha$, implying $F_n^{(\alpha)} \leq s_i$.
- Thus $F_n^{(\alpha)} \in [t_i, s_i]$ for any i and all large n, so
$$F_n^{(\alpha)} \to F^{(\alpha)}. \tag{4.1.4}$$

□

Example 4.1.2. Let $b_n^{(\alpha)}$ be the α^{th} quantile of $Binomial(n, \theta)$. By Central Limit Theorem, if $Y_n \sim Binomial(n, \theta)$, then
$$\frac{Y_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{\text{d}} N(0,1). \tag{4.1.5}$$
So
$$\frac{b_n^{(\alpha)} - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{\text{d}} z^{(\alpha)}, \tag{4.1.6}$$
the α^{th} quantile of $N(0,1)$ (i.e., $b_n^{(\alpha)} \approx n\theta + \sqrt{n\theta(1-\theta)} \cdot z^{(\alpha)}$ for large n).

Example 4.1.3. Let $(\mathcal{X}_n^2)^{(\alpha)}$ be the α^{th} quantile of $\mathcal{X}_n^2$. Recall $Y_n \sim \mathcal{X}_n^2$ when $Y_n = X_1^2 + X_2^2 + \cdots + X_n^2$, $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$. Then by Central Limit Theorem,
$$\frac{Y_n - n}{\sqrt{2n}} \xrightarrow{\text{d}} N(0,1). \tag{4.1.7}$$
So
$$\frac{(\mathcal{X}_n^2)^{(\alpha)} - n}{\sqrt{2n}} \xrightarrow{\text{d}} z^{(\alpha)}, \tag{4.1.8}$$
the α^{th} quantile of $N(0,1)$.(i.e., $(\mathcal{X}_n^2)^{(\alpha)} \approx n + \sqrt{2n} \cdot z^{(\alpha)}$ for large n).

Theorem 4.1.3 (Weak Law of Large Numbers)**.** If $X_1, X_2, \cdots, X_n \in \mathbb{R}^k$ are i.i.d. with mean $\mathbb{E}[X_i] = \mu \in \mathbb{R}^k$, then
$$\bar{X} \xrightarrow{\text{P}} \mu \text{ as } n \to \infty. \tag{4.1.9}$$

Definition 4.1.3 (Converge in probability)**.** Random vectors $Y_n \in \mathbb{R}^k$ *converge in probability* to $Y \in \mathbb{R}^k$ if for any $\epsilon > 0$,
$$\mathbb{P}[\|Y_n - Y\| > \epsilon] \to 0 \text{ as } n \to \infty. \tag{4.1.10}$$
We'll write $Y_n \xrightarrow{\text{P}} Y$.

Note 4.1.2. For a constant vector $c \in \mathbb{R}^k$,

$$Y_n \xrightarrow{\text{P}} c \iff Y_n \xrightarrow{\text{d}} c. \tag{4.1.11}$$

Example 4.1.4. Suppose $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} p_\theta$ and there is a function $f(X)$ such taht $f(X_i)$ is unbiased for $g(\theta)$ (i.e., $\mathbb{E}_\theta[f(X_i)] = g(\theta)$). Then the estimator

$$T(X_1, X_2, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

satisfies

$$T(X_1, X_2, \cdots, X_n) \xrightarrow{\text{P}} g(\theta) \text{ as } n \to \infty. \tag{4.1.12}$$

Furthermore,

$$\sqrt{n}(T(X_1, X_2, \cdots, X_n) - g(\theta)) \xrightarrow{\text{d}} N(0, Var_\theta[f(X_i)]). \tag{4.1.13}$$

Definition 4.1.4 (Asymptotically consistent and normal)**.** Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$. A sequence of estimators $T(X) =\equiv T(X_1, X_2, \cdots, X_n)$ is *asymptotically consistent* for estimating $g(\theta)$ if under every $\theta \in \Omega$,

$$T(X) \xrightarrow{\text{P}} g(\theta) \text{ as } n \to \infty. \tag{4.1.14}$$

$T(X)$ is *asymptotically normal* if

$$\sqrt{n}(T(X) - g(\theta)) \xrightarrow{\text{d}} N(0, V(\theta)). \tag{4.1.15}$$

for some limit variance $V(\theta) < \infty$.

Example 4.1.5. Consider $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$. Test

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

The likelihood ratio is

$$L(X) = L(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}.$$

Let

$$T(X) = \frac{1}{n}\log L(X) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}.$$

By Neyman-Pearson Lemma, the most powerful test rejects H_0 for large $T(X)$.

– Under $\theta = \theta_0$:

$$\begin{aligned} T(X) \xrightarrow{\text{P}} & \mathbb{E}_{\theta_0}[\log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}] \\ &= \int \log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} \cdot p_{\theta_0}(X_i)\mathrm{d}x \\ &= -\int \log \frac{p_{\theta_0}(X_i)}{p_{\theta_1}(X_i)} \cdot p_{\theta_0}(X_i)\mathrm{d}x \\ &= -D_{KL}(p_{\theta_0}||p_{\theta_1}). \end{aligned} \tag{4.1.16}$$

– Under $\theta = \theta_1$:

$$\begin{aligned} T(X) \xrightarrow{\mathrm{P}} & \mathbb{E}_{\theta_1}\left[\log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}\right] \\ =& \int \log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} \cdot p_{\theta_1}(X_i)\mathrm{d}x \\ =& D_{KL}(p_{\theta_1}||p_{\theta_0}). \end{aligned} \tag{4.1.17}$$

Note 4.1.3. When $p_{\theta_1} \neq p_{\theta_0}$, $D_{KL}(p_{\theta_0}||p_{\theta_1}), D_{KL}(p_{\theta_1}||p_{\theta_0}) > 0$ by Jensen's Inequality. So the test which rejects H_0 when $T(X) > 0$ is "asymptotically perfect": As $n \to \infty$

– Under $H_0 : \mathbb{P}_{\theta_0}[\text{Type I error}] \to 0$.

– Under $H_1 : \mathbb{P}_{\theta_1}[\text{Type II error}] \to 0$.

Question 4.1.1. How to understand asymptotic behavior of more complicated statistics?

Answer 4.1.1.

– Slutsky's Lemma.

– Continuous mapping theorem.

– Delta method (Taylor expansions).

4.2 Slutsky's Lemma

Theorem 4.2.1 (Slutsky's Lemma)**.** If $X_n \xrightarrow{\mathrm{d}} X$ and $Y_n \xrightarrow{\mathrm{P}} c$ (a constant), then

- $X_n + Y_n \xrightarrow{\mathrm{d}} X + c$.
- $X_n \cdot Y_n \xrightarrow{\mathrm{d}} cX$.

Proof. Let $t = x + c$ be a point where the CDF F of $X + c$ is continuous. Then the CDF of X is also continuous at x. We need to show

$$\mathbb{P}[X_n + Y_n \leq t] \to F(t) \text{ (i.e., } \mathbb{P}[X_n + Y_n \leq t] \to \mathbb{P}[X \leq x]). \tag{4.2.1}$$

– Take $\epsilon > 0$ such that the CDF of X is continuous at $x + \epsilon$. Then as $n \to \infty$,

$$\mathbb{P}[X_n + Y_n \leq x + c] \leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[Y_n \leq c - \epsilon] \to \mathbb{P}[X \leq x + \epsilon] + 0.$$

So for large n,

$$\mathbb{P}[X_n + Y_n \leq x + c] \leq \mathbb{P}[X \leq x + \epsilon] + \epsilon.$$

– Take $\epsilon > 0$ such that the CDF of X is continuous at $x - \epsilon$. Since

$$\mathbb{P}[X_n \leq x - \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x + c] + \mathbb{P}[Y_n \geq c + \epsilon]$$

Then as $n \to \infty$,

$$\mathbb{P}[X_n + Y_n \leq x + c] \geq \mathbb{P}[X_n \leq x - \epsilon] - \mathbb{P}[Y_n \geq c + \epsilon] \to \mathbb{P}[X \leq x - \epsilon] - 0.$$

So for large n,

$$\mathbb{P}[X_n + Y_n \leq x + c] \geq \mathbb{P}[X \leq x - \epsilon] - \epsilon.$$

– Take $\epsilon \to 0$

$$\mathbb{P}[X_n + Y_n \le t] \to \mathbb{P}[X \le x]. \tag{4.2.2}$$

The proof for $X_n \cdot Y_n \xrightarrow{\text{d}} cX$ is similar. □

Corollary 4.2.1. If $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\text{P}} c$ (a constant), and the CDF of X is continuous at c, then

$$\mathbb{P}[X_n \le Y_n] \to \mathbb{P}[X \le c]. \tag{4.2.3}$$

Proof. By Slutsky's Lemma, $X_n - Y_n \xrightarrow{\text{d}} X - c$. The cumulative dsitribution function of $X - c$ is continuous at 0. So

$$\mathbb{P}[X_n \le Y_n] = \mathbb{P}[X_n - Y_n \le 0] \to \mathbb{P}[X - c \le 0] = \mathbb{P}[X \le c]. \tag{4.2.4}$$

□

4.3 Continuous Mapping Theorem

Theorem 4.3.1 (Continuous Mapping Theorem)**.** Suppose $g : \mathbb{R}^k \to \mathbb{R}^s$ is continuous.

– If $X_n \xrightarrow{\text{d}} X \in \mathbb{R}^k$, then $g(X_n) \xrightarrow{\text{d}} g(X)$.

– If $X_n \xrightarrow{\text{P}} X \in \mathbb{R}^k$, then $g(X_n) \xrightarrow{\text{P}} g(X)$.

More generally, this holds as long as g is continuous on some set $C \subset \mathbb{R}^k$ where $\mathbb{P}[X \in C] = 1$.

Example 4.3.1 (Sample variance, t-statistic)**.** Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

$$T = \frac{\sqrt{n}\bar{X}}{S}.$$

For finite n,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

$$S^2 \sim \frac{\sigma^2}{n-1} \cdot \mathcal{X}^2_{n-1},$$

$$T \sim t_{n-1} \text{ when } \mu = 0.$$

For $n \to \infty$,

– For $\bar{X}$, by Central Limit Theorem

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{\text{d}} N(\mu, \sigma^2). \tag{4.3.1}$$

– For S^2, note that

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \mu)^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\mu - \bar{X}) + n(\mu - \bar{X})^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.\end{aligned}$$

So

$$S^2 = \frac{n}{n-1} \cdot \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{n-1} \cdot (\bar{X} - \mu)^2.$$

Here,

- $\frac{n}{n-1} \to 1$.
- $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{\text{P}} \sigma^2$ by Weak Law of Large Numbers.
- $(\bar{X} - \mu) \xrightarrow{\text{P}} 0$ by Weak Law of Large Numbers, so $(\bar{X} - \mu)^2 \xrightarrow{\text{P}} 0$ by Continuous Mapping Theorem.

Then the Slutsky's Lemma implies

$$S^2 \xrightarrow{\text{P}} 1 \cdot \sigma^2 - 1 \cdot 0 = \sigma^2. \tag{4.3.2}$$

– For T, when $\mu = 0$, by Slutsky's Lemma and Continuous Mapping Theorem

$$T = \frac{\sqrt{n}\bar{X}}{S} \xrightarrow{\text{d}} N(0, \sigma^2) \cdot \frac{1}{\sigma} = N(0, 1). \tag{4.3.3}$$

Also, by Convergence of Quantiles,

$$t_{n-1}^{(1-\alpha)} \to z^{(1-\alpha)}. \tag{4.3.4}$$

So

$$\mathbb{P}_{\mu=0}[T > t_{n-1}^{(1-\alpha)}] \to \mathbb{P}[Z > z^{(1-\alpha)}], \text{ where } Z \sim N(0, 1). \tag{4.3.5}$$

Thus, as $n \to \infty$, the Type I error of t-test $\to \alpha$ (i.e., *asymptotically level* α). This *didn't* use $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, only that $\mathbb{E}[X_i] = 0, Var[X_i] = \sigma^2$.

Question 4.3.1. What about the second-order behavior of S^2?

Answer 4.3.1.

$$\begin{aligned} W &= \sqrt{n}(S^2 - \sigma^2) \\ &= \sqrt{n}\left(\frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{n}{n-1}\cdot(\bar{X}-\mu)^2 - \sigma^2\right) \\ &= \sqrt{n}\left(\frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu)^2-\sigma^2] - \frac{n}{n-1}\cdot(\bar{X}-\mu)^2 + \frac{1}{n-1}\cdot\sigma^2\right) \\ &= \frac{n}{n-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(X_i-\mu)^2-\sigma^2\right] - \frac{\sqrt{n}}{n-1}\cdot[\sqrt{n}(\bar{X}-\mu)]^2 + \frac{\sqrt{n}}{n-1}\cdot\sigma^2. \end{aligned} \tag{4.3.6}$$

Here,

- $\frac{n}{n-1} \to 1$.
- $\frac{\sqrt{n}}{n-1} \to 0$.
- $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(X_i-\mu)^2-\sigma^2] \xrightarrow{\text{d}} N(0, Var[(X_i-\mu)^2])$.
- $\sqrt{n}(\bar{X}-\mu) \xrightarrow{\text{d}} N(0,1)$.

By slutsky's Lemma, $W \to N(0, Var[(X_i-\mu)^2])$.

If $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then $Var[(X_i-\mu)^2] = 2\sigma^4$. In general, it depends on 4^{th} moment of X_i.

4.4 Delta Method

Theorem 4.4.1 (Delta Method)**.**

(i) Suppose random variables $X_n \in \mathbb{R}$ satisfy $\tau_n(X_n - \mu) \xrightarrow{\text{d}} Z$ for some $\mu \in \mathbb{R}$, sequence $\tau_n \to \infty$, and limit distribution $Z \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable at μ. Then

$$\tau_n(g(X_n) - g(\mu)) \xrightarrow{\text{d}} g'(\mu)\cdot Z. \tag{4.4.1}$$

(ii) Suppose random vectors $X_n \in \mathbb{R}^k$ satisfy $\tau_n(X_n - \mu) \xrightarrow{\text{d}} Z$ for some $\mu \in \mathbb{R}^k$, sequence $\tau_n \to \infty$, and limit distribution $Z \in \mathbb{R}^k$. Let $g : \mathbb{R}^k \to \mathbb{R}^s$ be differentiable at μ with derivative $Dg(\mu) \in \mathbb{R}^{s\times k}$. Then

$$\tau_n(g(X_n) - g(\mu)) \xrightarrow{\text{d}} (Dg(\mu))\cdot Z. \tag{4.4.2}$$

Proof.

(i) Since g is differentiable at μ,

$$g(X) = g(\mu) + g'(\mu) \cdot (X - \mu) + R(X - \mu)$$

where $\frac{R(X-\mu)}{X-\mu} \to 0$ as $X \to \mu$. Then for $X = X_n$,

$$\tau_n(g(X_n) - g(\mu)) = g'(\mu) \cdot \tau_n(X_n - \mu) + \tau_n \cdot R(X_n - \mu). \tag{4.4.3}$$

Define $h(X - \mu) = \frac{R(X-\mu)}{X-\mu}$, where $h(X - \mu) \to 0$ as $X \to \mu$. Then

$$\tau_n(g(X_n) - g(\mu)) = g'(\mu) \cdot \tau_n(X_n - \mu) + \tau_n(X_n - \mu) \cdot h(X_n - \mu). \tag{4.4.4}$$

Here,

- $\tau_n(X_n - \mu) \xrightarrow{\text{d}} Z$ by assumption.
- $X_n - \mu = \frac{1}{\tau_n} \cdot \tau_n(\tau_n - \mu) \xrightarrow{\text{P}} 0$ by Slutsky's Lemma.
- $h(X_n - \mu) \xrightarrow{\text{P}} 0$ by Continuous Mapping Theorem.
- $\tau_n(X_n - \mu) \cdot h(X_n - \mu) \xrightarrow{\text{P}} 0$ by Slutsky's Lemma.

So

$$\tau_n(g(X_n) - g(\mu)) \xrightarrow{\text{d}} g'(\mu) \cdot Z. \tag{4.4.5}$$

(ii) The proof for (ii) is analogous by using

$$g(X) = g(\mu) + (Dg(\mu)) \cdot (X - \mu) + R(X - \mu).$$

□

Corollary 4.4.1.

(i) If $\sqrt{n}(T_n - \mu) \xrightarrow{\text{d}} N(0, \sigma^2) \in \mathbb{R}$. Then

$$\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{\text{d}} N(0, g'(\mu)^2\sigma^2). \tag{4.4.6}$$

(ii) If $\sqrt{n}(T_n - \mu) \xrightarrow{\text{d}} N(0, \Sigma) \in \mathbb{R}^k$. Then

$$\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{\text{d}} N(0, D\Sigma D^T) \text{ where } D = Dg(\mu). \tag{4.4.7}$$

Example 4.4.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. Then by Central Limit Theorem, $\sqrt{n}(\bar{X} - \theta) \xrightarrow{\text{d}} N(0, \theta(1-\theta))$. We can estimate the variance $g(\theta) = \theta(1-\theta)$ by the *plug-in estimator* $\delta(X) = \bar{X}(1 - \bar{X})$. What's the behavior of $\delta(X)$ as $n \to \infty$?

(i) First-order: $\bar{X} \xrightarrow{\text{P}} \theta$. By Continuous Mapping Theorem, $\delta(X) \xrightarrow{\text{P}} \theta(1-\theta) = g(\theta)$ (i.e., $\delta(X)$ is consistent).

(ii) Second-order: $g'(\theta) = 1 - 2\theta$. By Delta Method, $\sqrt{n}(\delta(X) - g(\theta)) \xrightarrow{\text{d}} N(0, (1-2\theta)^2 \cdot \theta(1-\theta))$ (i.e., $\delta(X)$ is asymptotically normal).

Example 4.4.2. Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. with mean μ and variance σ^2. Consider $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2\right]$. Let $Y_i = (Y_i^{(1)}, Y_i^{(2)}) = \left(X_i - \mu, (X_i-\mu)^2 - \sigma^2\right)$. Then

$$Cov(Y_i) \equiv \Sigma \equiv \begin{pmatrix} \sigma^2 & \alpha \\ \alpha & \nu \end{pmatrix} \text{ where } \alpha = Cov[Y_i^{(1)}, Y_i^{(2)}],\ \nu = Var(Y_i^{(2)}).$$

By Multivariate Central Limit Theorem,

$$\sqrt{n}\cdot\bar{Y} \xrightarrow{\text{d}} N(0,\Sigma).$$

Note: $\frac{n-1}{n}S^2 - \sigma^2 = \bar{Y}^{(2)} - (\bar{Y}^{(1)})^2$ where $\bar{Y} = (\bar{Y}^{(2)}, \bar{Y}^{(1)})$.

Let $f(a,b) = b - a^2$. Then $Df(x,y) = (-2a, 1)$. By Delta Method,

$$\begin{aligned}\sqrt{n}\left(\frac{n-1}{n}S^2 - \sigma^2\right) &= \sqrt{n}\left(\bar{Y}^{(2)} - (\bar{Y}^{(1)})^2\right) \\ &= \sqrt{n}(f(\bar{Y}^{(2)}, \bar{Y}^{(1)}) - f(0,0)) \\ &\xrightarrow{\text{d}} N\left(0, Df(0,0)\cdot\Sigma\cdot Df(0,0)^T\right) = N(0,\nu).\end{aligned} \tag{4.4.8}$$

Then also

$$\sqrt{n}(S^2 - \sigma^2) = \sqrt{n}\left(\frac{n-1}{n}S^2 - \sigma^2\right) + \frac{1}{\sqrt{n}}S^2 \xrightarrow{\text{d}} N(0, v). \tag{4.4.9}$$

Definition 4.4.1 (Confidence interval)**.** For estimating $g(\theta)$ based on $X = (X_1, X_2, \cdots, X_n)$, a *confidence interval* $[L(X), U(X)]$ has coverage $1-\alpha$ if

$$\mathbb{P}_\theta[L(X) \le g(\theta) \le U(X)] \ge 1-\alpha \text{ for every } \theta \in \Omega. \tag{4.4.10}$$

It has asymptotic coverage $1-\alpha$ if

$$\lim_{n\to\infty} \mathbb{P}_\theta[L(X) \le g(\theta) \le U(X)] \ge 1-\alpha \text{ for every } \theta \in \Omega. \tag{4.4.11}$$

Note 4.4.1. $g(\theta)$ is fixed, $L(X), U(X)$ are the random quantities.

Proposition 4.4.1. Suppose $T_n = T_n(X_1, X_2, \cdots, X_n)$ be an asymptotically normal estimator for $g(\theta)$. So

$$\sqrt{n}(T_n - g(\theta)) \xrightarrow{\text{d}} n(0, V(\theta)) \text{ for some } v(\theta) > 0.$$

Suppose $V_n = V_n(X_1, X_2, \cdots, X_n)$ is asymptotically consistent for $V(\theta)$. So

$$V_n \xrightarrow{\text{P}} V(\theta).$$

Then an confidence interval with asymptotic coverage $1-\alpha$ for $g(\theta)$ is

$$\left[T_n - z^{(1-\frac{\alpha}{2})}\cdot\sqrt{\frac{V_n}{n}}, T_n + z^{(1-\frac{\alpha}{2})}\cdot\sqrt{\frac{V_n}{n}}\right] \tag{4.4.12}$$

Proof.

$$\mathbb{P}_\theta\left[T_n - z^{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{V_n}{n}} \le g(\theta) \le T_n + z^{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{V_n}{n}}\right]$$
$$= \mathbb{P}_\theta\left[-z^{(1-\frac{\alpha}{2})} \le \sqrt{\frac{n}{V_n}} \cdot (T_n - g(\theta)) \le z^{(1-\frac{\alpha}{2})}\right].$$

Since

- $\sqrt{n}(T_n - g(\theta)) \xrightarrow{\text{d}} N(0, V(\theta))$ by assumption.
- $\frac{1}{\sqrt{V_n}} \xrightarrow{\text{P}} \frac{1}{\sqrt{V(\theta)}}$ by assumption and Continuous Mapping Theorem.
- $\sqrt{\frac{n}{V_n}} \cdot (T_n - g(\theta)) \xrightarrow{\text{d}} N(0,1)$ by Slutsky's Lemma.

This probability converges to $\mathbb{P}[-z^{(1-\frac{\alpha}{2})} \le Z \le z^{(1-\frac{\alpha}{2})}] = 1 - \alpha$. □

Example 4.4.3. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. Then

- $\sqrt{n}(\bar{X} - \theta) \xrightarrow{\text{d}} N(0, \theta(1-\theta))$.
- $V_n = \bar{X}(1-\bar{X}) \xrightarrow{\text{P}} \theta(1-\theta)$.

Thus an confidence interval with asymptotic coverage $1-\alpha$ for $g(\theta)$ is

$$\left[\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z^{(1-\frac{\alpha}{2})}, \bar{X} + \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z^{(1-\frac{\alpha}{2})}\right]. \tag{4.4.13}$$

This is the Wald interval for θ.

Note 4.4.2. This type of interval is asymptotic and may not have coverage $1-\alpha$. Its accuracy depends on two approximations:

(i) $\sqrt{n}(T_n - g(\theta)) \approx N(0, V(\theta))$.

(ii) $V_n \approx V(\theta)$.

We can eliminate the second approximation by a *variance-stabilizing transformation*: Suppose $\sqrt{n}(T_n - g(\theta)) \xrightarrow{\text{d}} N(0, V(\theta))$ and there is a finite function $f(\theta)$ where $f'(\theta) = \frac{1}{V(\theta)} > 0$. Then by Delta Method,

$$\sqrt{n}(f(T_n) - f(\theta)) \xrightarrow{\text{d}} N(0,1). \tag{4.4.14}$$

The variance is no longer depends on θ. An asymptotic confidence interval for $f(\theta)$ is

$$I = \left[f(T_n) - \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})}, f(T_n) + \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})}\right]. \tag{4.4.15}$$

Note that $f'(\theta) = \frac{1}{V(\theta)} > 0$, so f is increasing and invertible. Thus an confidence interval with asymptotic coverage $1-\alpha$ for θ is

$$f^{-1}(I) = \left[f^{-1}(f(T_n) - \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})}), f^{-1}(f(T_n) + \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})})\right]. \tag{4.4.16}$$

This is no longer symmetric around T_n, but might have more accurate coverage than $\left[T_n - z^{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{V_n}{n}}, T_n + z^{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{V_n}{n}}\right]$ for finite n.

Example 4.4.4. Suppose $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_{YY} \end{pmatrix}$. To estimate the correlation $\rho = \frac{\Sigma_{XY}}{\sqrt{\Sigma_{XX}\Sigma_{YY}}}$, consider the sample correlation $\hat{\rho} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$ where $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Let $W_i = (W_i^{(1)}, W_i^{(2)}, W_i^{(3)}, W_i^{(4)}, W_i^{(5)}) = (X_i, Y_i, X_i^2, Y_i^2, X_i Y_i)$, then

$$\begin{aligned} \hat{\rho_n} &= \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2\right)}} \\ &= \frac{\bar{W}^{(5)} - \bar{W}^{(1)}\bar{W}^{(2)}}{(\bar{W}^{(3)} - (\bar{W}^{(1)})^2)(\bar{W}^{(4)} - (\bar{W}^{(2)})^2)}. \end{aligned}$$

By Central Limit Theorem,

$$\sqrt{n}(\bar{W} - (0, 0, \Sigma_{XX}, \Sigma_{YY}, \Sigma_{XY})) \xrightarrow{\text{d}} N(0, V) \text{ for some } V \in \mathbb{R}^{5\times 5}.$$

By Delta Method,

$$\sqrt{n}(\hat{\rho_n} - \rho) \xrightarrow{\text{d}} N(0, (1-\rho)^2). \tag{4.4.17}$$

The "usual" confidence interval with asymptotic coverage $1 - \alpha$ is

$$\left[\hat{\rho_n} - \sqrt{\frac{(1 - \hat{\rho_n}^2)^2}{n}} \cdot Z^{(1-\frac{\alpha}{2})}, \hat{\rho_n} + \sqrt{\frac{(1 - \hat{\rho_n}^2)^2}{n}} \cdot Z^{(1-\frac{\alpha}{2})}\right] \tag{4.4.18}$$

Consider $f(\rho) = \tanh^{-1}(\rho)$, then $f'(\rho) = \frac{1}{1-\rho^2}$. By Delta Method,

$$\sqrt{n}(f(\hat{\rho_n} - f(\rho))) \xrightarrow{\text{d}} N(0, 1) \tag{4.4.19}$$

So a confidence interval with asymptotic coverage $1 - \alpha$ for $\tanh^{-1}(\rho)$ is

$$\left[\tanh^{-1}(\hat{\rho_n}) - \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})}, \tanh^{-1}(\hat{\rho_n}) + \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})}\right]. \tag{4.4.20}$$

So a confidence interval with asymptotic coverage $1 - \alpha$ for ρ is

$$[\tanh(\tanh^{-1}(\hat{\rho_n}) - \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})}), \tanh(\tanh^{-1}(\hat{\rho_n}) + \frac{1}{\sqrt{n}} \cdot z^{(1-\frac{\alpha}{2})})]. \tag{4.4.21}$$

For finite n, it has better coverage than the "usual" confidence interval.

Question 4.4.1. What happens in Delta Method if $g'(\theta) = 0$?

Answer 4.4.1. $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{\text{d}} N(0, 0)$ (i.e., $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{\text{P}} 0$). So we need to do Taylor expansion to the second order.

Example 4.4.5. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Bernoulli(\theta)$. By Example 4.4.1 $\sqrt{n}(\bar{X}(1-\bar{X}) - \theta(1-\theta)) \xrightarrow{\text{d}} N(0, (1-2\theta)^2 \cdot \theta(1-\theta))$. At $\theta = \frac{1}{2}$, $g'(\theta) = 0$. So this limit is degenerate: $\sqrt{n}\left(\bar{X}(1-\bar{X}) - \frac{1}{4}\right) \xrightarrow{\text{P}} 0$. Let $g(X) = X(1-X)$, then

$$g(X) = g(\theta) + g'(\theta) \cdot (X-\theta) + \frac{1}{2} g''(\theta) \cdot (x-\theta)^2 + R((X-\mu)^2)$$

where $\frac{R((X-\theta)^2)}{(X-\theta)^2} \to 0$ as $X \to \theta$.

Then for $X = \bar{X}$,

$$n(g(\bar{X}) - g(\theta)) = g'(\theta) \cdot n(\bar{X}-\theta) + \frac{1}{2} g''(\theta) \cdot n(\bar{X}-\theta)^2 + R((\bar{X}-\theta)^2). \quad (4.4.22)$$

Define $h(X-\theta) = \frac{R((\bar{X}-\theta)^2)}{(X-\theta)^2}$, where $h(X-\theta) \to 0$ as $X \to \theta$. Then

$$n(g(\bar{X}) - g(\theta)) = g'(\theta) \cdot n(\bar{X}-\theta) + \frac{1}{2} g''(\theta) \cdot n(\bar{X}-\theta)^2 + n(\bar{X}-\theta)^2 \cdot h(\bar{X}-\theta). \quad (4.4.23)$$

Here,

- $\sqrt{n}(\bar{X}-\theta) \xrightarrow{\text{d}} N(0, \theta(1-\theta))$ by Central Limit Theorem.
- $n(\bar{X}-\theta)^2 \xrightarrow{\text{d}} \theta(1-\theta)\mathcal{X}_1^2$ by Continuous Mapping Theorem.
- $n(\bar{X}-\theta)^2 \cdot h(\bar{X}-\theta)$ by Slutsky's Theorem.

So at $\theta = \frac{1}{2}$, $g'(\theta) = 0$,

$$\begin{aligned} n(g(\bar{X}) - g(\theta)) &\xrightarrow{\text{d}} \frac{1}{2} g''(\theta) \cdot \theta(1-\theta)\mathcal{X}_1^2. \\ n\left(\bar{X}(1-\bar{X}) - \frac{1}{4}\right) &\xrightarrow{\text{d}} -\frac{1}{4}\mathcal{X}_1^2. \end{aligned} \quad (4.4.24)$$

The left scaling is now n, rather than $\sqrt{n}$.

5 Pointwise Asymptotics of MLE

5.1 Maximum Likelihood Estimation (MLE)

Definition 5.1.1 (Likelihood function)**.** Given data $X = (X_1, X_2, \cdots, X_n) \sim p_\theta(x_1, x_2, \cdots, x_n)$, the density $p_\theta(x_1, x_2, \cdots, x_n)$ viewed as a function of θ is called the *likelihood function*. We denote by $l(\theta|x) = \log p_\theta(x_1, x_2, \cdots, x_n)$ the *log-likelihood*.

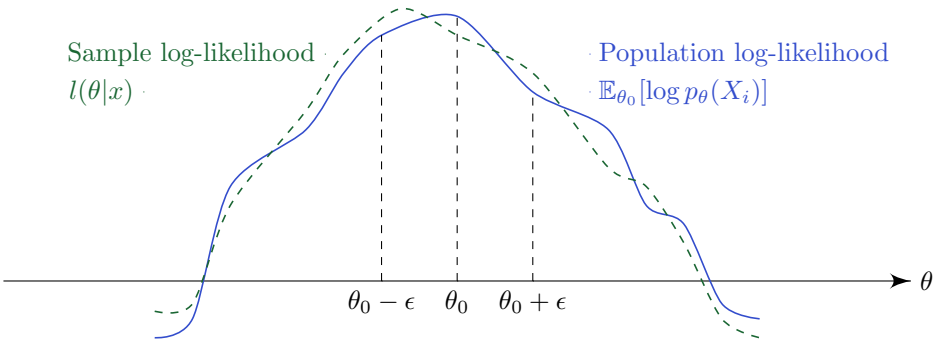
Note 5.1.1. If $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, then $p_\theta(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p_\theta(x_i)$, $l(\theta|x) = \sum_{i=1}^{n} \log p_\theta(x_i)$. By Weak Law of Large Numbers, if the true parameter is θ_0, then as $n \to \infty$

$$\frac{1}{n} \cdot l(\theta|x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \log p_\theta(x_i) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[\log p_\theta(X_i)]. \tag{5.1.1}$$

$$\text{Sample log-likelihood} \xrightarrow{\text{P}} \text{Population log-likelihood.}$$

The population log-likelihood is maximized over θ at the true parameter $\theta = \theta_0$:

$$\begin{aligned}
\mathbb{E}_{\theta_0}[\log p_{\theta_0}(X_i)] - \mathbb{E}_{\theta_0}[\log p_\theta(X_i)] &= \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right] \\
&= -\mathbb{E}_{\theta_0}\left[\log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right] \\
&\geq -\log \mathbb{E}_{\theta_0}\left[\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right] \quad \text{By Jensen's Inequality} \\
&= -\log \int \frac{p_\theta(x_i)}{p_{\theta_0}(x_i)} \cdot p_{\theta_0}(x_i) \mathrm{d}x \\
&= -\log \int p_\theta(x_i) \mathrm{d}x \\
&= 0.
\end{aligned} \tag{5.1.2}$$



Definition 5.1.2 (MLE)**.** The *maximum likelihood estimator* (MLE) of θ is

$$\hat{\theta} = \underset{\theta \in \Omega}{\operatorname{argmax}}\, l(\theta|x). \tag{5.1.3}$$

If we want to estimate $g(\theta)$, we typically use the "plug-in" MLE $g(\hat{\theta})$.

Example 5.1.1. (MLE not finite-n "optimal") Suppsoe $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \nu)$, $\theta = (\mu, \nu)$.

$$l(\mu, \nu | x) = \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi\nu) - \frac{(x_i - \mu)^2}{2\nu} \right].$$

Check: $l(\mu, \nu|x) \to -\infty$ as $\mu \to \pm\infty,\ \nu \to \{0, \infty\}$. So maximum occurs at a point (μ, ν) where

$$0 = \frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\nu}$$

$$0 = \frac{\partial l}{\partial \nu} = -\frac{n}{2\nu} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\nu^2}$$

The MLE is $(\hat{\mu}, \hat{\nu}) = (\bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2) = (\bar{X}, \frac{1}{n} S^2)$.
The unbiased estimate of ν is $\frac{1}{n-1} S^2$.
Under the squared-error loss $(\hat{\nu} - \nu)^2$, both are inadmissible and dominated by $\hat{\nu} = \frac{1}{n+1} S^2$. (Check: $\mathbb{E}[(cS^2 - \nu)^2] = (c(n-1) - 1)^2 \cdot \nu^2 + c^2 \cdot 2(n-1) \cdot \nu^2$, minimized at $c = \frac{1}{n+1}$.)

5.2 Asymptotic Consistency of MLE

Recall: An estimator $\hat{\theta}_n$ is consistent for θ at the true parameter θ_0 if for any $\epsilon > 0$,

$$\mathbb{P}_{\theta_0}[\|\hat{\theta}_n - \theta_0\| \le \epsilon] \to 1 \text{ as } n \to \infty. \tag{5.2.1}$$

Theorem 5.2.1. Consider $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$ for $\theta \in \Omega \subseteq \mathbb{R}$. Let $l(\theta|x_i) = \log p_\theta(x_i),\ l(\theta|x) = \sum_{i=1}^{n} l(\theta|x_i)$. Suppose Ω contains an open interval ω around the true parameter θ_0, and $l(\theta|x_i)$ is continuous at all $\theta \in \omega$. Then for any $\epsilon > 0$, as $n \to \infty$,

$$\mathbb{P}_{\theta_0}[\text{ there exists a local maximizer } \hat{\theta}_n \text{ of } l(\theta|x) : \|\hat{\theta}_n - \theta_0\|] = 1. \tag{5.2.2}$$

Proof. Pick any $\epsilon > 0$ such that $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset \omega$. Compare $\frac{1}{n} \cdot l(\theta|x)$ at $\theta_0, \theta_0 - \epsilon, \theta_0 + \epsilon$. By Weak Law of Large Numbers,

- $\frac{1}{n} \cdot l(\theta_0|x) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[l(\theta_0|X_i)]$.
- $\frac{1}{n} \cdot l(\theta_0 - \epsilon|x) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[l(\theta_0 - \epsilon|X_i)]$.
- $\frac{1}{n} \cdot l(\theta_0 + \epsilon|x) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[l(\theta_0 + \epsilon|X_i)]$.

By (5.1.2),

$$\mathbb{E}_{\theta_0}[l(\theta_0|X_i)] > \max(\mathbb{E}_{\theta_0 - \epsilon}[l(\theta_0|X_i)], \mathbb{E}_{\theta_0 + \epsilon}[l(\theta_0|X_i)]).$$

Then with probability approaching 1 as $n \to \infty$,

$$\frac{1}{n} \cdot l(\theta_0|x) > \max\left(\frac{1}{n} \cdot l(\theta_0 - \epsilon|x), \frac{1}{n} \cdot l(\theta_0 - \epsilon|x) \right).$$

Since $l(\theta|x)$ is continuous on ω, then there must exist a local maximum $\hat{\theta}_n$ of $l(\theta|x)$ in $(\theta_0 - \epsilon, \theta_0 + \epsilon)$ when the above occurs. And at $\hat{\theta}_n$, we have $l'(\hat{\theta}_n|x) = 0$. $\square$

Corollary 5.2.1. Suppose the MLE $\hat{\theta}_n$ is the unique local maximzer of $l(\theta|x)$ with probability approaching 1 as $n \to \infty$ under the true parameter θ_0. Then $\hat{\theta}_n$ is asymptotically consistent for θ at θ_0. (e.g., If $l(\theta|x)$ is strictly concave in θ.)

Example 5.2.1. Consider a 1-parameter, exponential family model

$$p(x_i|\eta) = e^{\eta T(x_i) - A(\eta)} h(x_i).$$

Then

$$\begin{aligned} l(\eta|x) &= \sum_{i=1}^{n} (\eta T(x_i) - A(\eta) + \log h(x_i)). \\ l'(\eta|x) &= \sum_{i=1}^{n} T(x_i) - nA'(\eta). \\ l''(\eta|x) &= -nA''(\eta). \end{aligned}$$

Recall Remark 1.4.4 and (1.4.13),

$$A'(\eta) = \mathbb{E}_\eta[T_i(X)], \ A''(\eta) = Var_\eta[T_i(X)].$$

So $l''(\eta|x) = -nA''(\eta) < 0$, $l(\eta|x)$ is concave in η. Then

$$0 = l'(\eta|x) \iff \frac{1}{n} \sum_{i=1}^{n} T(x_i) = A'(\eta) = \mathbb{E}_\eta[T(X_i)].$$

has at most one root $\hat{\eta}$. When $\hat{\eta}$ exists, it is the unique local maximzer. By Corollary 5.2.1, it must be the global MLE.

Theorem 5.2.2. Consider $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$ for $\theta \in \Omega \subseteq \mathbb{R}^k$. Let $l(\theta|x_i) = \log p_\theta(x_i)$, $l(\theta|x) = \sum_{i=1}^{n} l(\theta|x_i)$. Suppose Ω contains an open interval ω around the true parameter θ_0 such that

(i) $l(\theta|x_i)$ is differentiable at every $\theta \in \omega$.

(ii) There is a function $m(x_i)$ such that $\|\nabla l(\theta|x_i)\| \le m(x_i)$ for all $\theta \in \omega$ and $\mathbb{E}_{\theta_0}[m(X_i)] < \infty$.

Then for any $\epsilon > 0$, as $n \to \infty$,

$$\mathbb{P}_{\theta_0}[\text{ there exists a local maximizer } \hat{\theta}_n \text{ of } l(\theta|x) : \|\hat{\theta}_n - \theta_0\|] = 1. \tag{5.2.3}$$

Proof. Let S be the sphere of radius ϵ around θ_0. Take $\epsilon > 0$ such that $S \subset \omega$. It suffices to show with probabbility approaching 1 as $n \to \infty$,

$$\frac{1}{n} \cdot l(\theta_0|x) > \sup_{\theta \in S} \frac{1}{n} \cdot l(\theta|x). \tag{5.2.4}$$

Since this would imply that $l(\theta|x)$ has a local maximum $\hat{\theta}_n$ inside S.

– Analysis of populatioin log-likelihood $\mathbb{E}_{\theta_0}[l(\theta|X_i)]$:
Note first that (ii) implies $\theta \to \mathbb{E}_{\theta_0}[l(\theta|X_i)]$ is continuous over $\theta \in \omega$:
For any $\theta, \theta' \in \omega$, let $\theta_t = t\theta' + (1-t)\theta$ for any $t \in [0,1]$. Then

$$\begin{aligned}|l(\theta'|X_i) - l(\theta|X_i)| &= \left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} l(\theta_t|x_i)\mathrm{d}t\right| \\ &= \left|\int_0^1 (\theta'-\theta)^T \nabla l(\theta_t|x_i)\mathrm{d}t\right| \\ &\le \|\theta'-\theta\| \cdot \sup_{\tilde{\theta}\in\omega} \|\nabla l(\tilde{\theta}|x_i)\|.\end{aligned}$$

So

$$\begin{aligned}\lim_{\theta'\to\theta} |\mathbb{E}_{\theta_0}[l(\theta'|X_i)] - \mathbb{E}_{\theta_0}[l(\theta|X_i)]| &\le \lim_{\theta'\to\theta} \mathbb{E}_{\theta_0}[|l(\theta'|X_i) - l(\theta|X_i)|] \\ &\le \lim_{\theta'\to\theta} \mathbb{E}_{\theta_0}[\|\theta'-\theta\| \cdot \sup_{\tilde{\theta}\in\omega} \|\nabla l(\tilde{\theta}|x_i)\|] \\ &\le \lim_{\theta'\to\theta} \|\theta'-\theta\| \cdot \mathbb{E}_{\theta_0}[m(X_i)] = 0.\end{aligned}$$

As $\theta' \to \theta$, $\mathbb{E}_{\theta_0}[l(\theta'|X_i)] \to \mathbb{E}_{\theta_0}[l(\theta|X_i)]$ (i.e., $\theta \to \mathbb{E}_{\theta_0}[l(\theta|X_i)]$ continuous).
Since S is compact, there exists some $\theta_* \in S$ where

$$\begin{aligned}&\sup_{\theta\in S} E_{\theta_0}[l(\theta|x_i)] = E_{\theta_0}[l(\theta_*|x_i)]. \\ &E_{\theta_0}[l(\theta_*|x_i)] < E_{\theta_0}[l(\theta_0|x_i)].\end{aligned}$$

Then

$$\delta \equiv \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] - \sup_{\theta\in S} \mathbb{E}_{\theta_0}[l(\theta|X_i)] = \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] - \mathbb{E}_{\theta_0}[l(\theta_*|X_i)] > 0. \tag{5.2.5}$$

– Analysis of sample log-likelihood $\frac{1}{n}l(\theta|x)$:
Let $N \subset S$ be a finite set, such that for each $\theta \in S$, there is $\mu(\theta) \in N$ with $\|\theta - \mu(\theta)\| < \eta$. Consider an event ε_n where

(a) $\left|\frac{1}{n}\cdot l(\theta_0|x) - \mathbb{E}_{\theta_0}[l(\theta_0|X_i)]\right| < \frac{\delta}{3}$.
(b) $\left|\frac{1}{n}\cdot l(\mu|x) - \mathbb{E}_{\theta_0}[l(\mu|X_i)]\right| < \frac{\delta}{3}$ for all $\mu \in N$.
(c) $\left|\frac{1}{n}\cdot l(\theta|x) - \frac{1}{n}\cdot l(\mu(\theta)|x)\right| < \frac{\delta}{3}$ for all $\theta \in S$.

Now we are going to show $\mathbb{P}_{\theta_0}[\varepsilon_n] \to 1$ as $n \to \infty$.

(a) By Weak Law of Large Numbers, as $n \to infty$,

$$\mathbb{P}_{\theta_0}\left[\left|\frac{1}{n}\cdot l(\theta_0|x) - \mathbb{E}_{\theta_0}[l(\theta_0|X_i)]\right| \ge \frac{\delta}{3}\right] \to 0.$$

(b) By Weak Law of Large Numbers, as $n \to infty$,

$$\begin{aligned}&\mathbb{P}_{\theta_0}\left[\text{ there exists } \mu \in N \text{ such that } \left|\frac{1}{n}\cdot l(\mu|x) - \mathbb{E}_{\theta_0}[l(\mu|X_i)]\right| \ge \frac{\delta}{3}\right] \\ &\le \sum_{\mu\in N} \mathbb{P}_{\theta_0}\left[\left|\frac{1}{n}\cdot l(\mu|x) - \mathbb{E}_{\theta_0}[l(\mu|X_i)]\right| \ge \frac{\delta}{3}\right] \to 0\end{aligned}$$

(c) By assumption (ii), as $n \to \infty$,

$$\begin{aligned}
&\mathbb{P}_{\theta_0}\left[\sup_{\theta\in S}\left|\frac{1}{n}\cdot l(\theta|x) - \frac{1}{n}\cdot l(\mu(\theta|x))\right| \geq \frac{\delta}{3}\right] \\
&\leq \mathbb{P}_{\theta_0}\left[\sup_{\theta\in S}\|\theta-\mu(\theta)\|\cdot\|\nabla\frac{1}{n}l(\theta_t|x)\| \geq \frac{\delta}{3}\right] \text{ where } \theta_t = t\mu(\theta)+(1-t)\theta \\
&\leq \mathbb{P}_{\theta_0}\left[\sup_{\tilde{\theta}\in S}\|\nabla\frac{1}{n}l(\tilde{\theta}|x)\| \geq \frac{\delta}{3\eta}\right] = \mathbb{P}_{\theta_0}\left[\sup_{\tilde{\theta}\in S}\|\frac{1}{n}\sum_{i=1}^{n}\nabla l(\tilde{\theta}|x_i)\| \geq \frac{\delta}{3\eta}\right] \\
&\leq \mathbb{P}_{\theta_0}\left[\frac{1}{n}\sum_{i=1}^{n}m(X_i) \geq \frac{\delta}{3\eta}\right] \\
&\leq \frac{3\eta}{\delta}\cdot\mathbb{E}_{\theta_0}\left[\frac{1}{n}\sum_{i=1}^{n}m(X_i)\right] \text{ by Markov's Inequality} \\
&= \frac{3\eta}{\delta}\cdot\mathbb{E}_{\theta_0}[m(X_i)].
\end{aligned}$$

Here $\eta > 0$ is arbitary, take $\eta \to$, then $\mathbb{P}_{\theta_0}[(c)^c] \to 0$.

Therefore, as $n \to 0$,

$$\mathbb{P}_{\theta_0}[\varepsilon_n^c] \leq \mathbb{P}_{\theta_0}[(a)^c] + \mathbb{P}_{\theta_0}[(b)^c] + \mathbb{P}_{\theta_0}[(c)^c] \to 0.$$

So with probability approaching 1 as $n \to \infty$, ε_n holds. Then $\forall\theta \in S$

$$\begin{aligned}
\frac{1}{n}\cdot l(\theta|x) &< \frac{1}{n}l(\mu(\theta)|x) + \frac{\delta}{3} \\
&< \mathbb{E}_{\theta_0}[l(\mu(\theta)|X_i)] + \frac{\delta}{3} + \frac{\delta}{3} \\
&< \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] - \frac{\delta}{3} \\
&< \frac{1}{n}\cdot l(\theta_0|x).
\end{aligned} \tag{5.2.6}$$

□

Example 5.2.2. Let $p(x_i|\eta) = \exp(\sum_{j=1}^{k}\eta_j T_j(x_i) - A(\eta))h(x_i)$ be a k-parameter exponential family, η_0 in the interior of Ξ. Then

$$\begin{aligned}
&l(\eta|x) = \sum_{i=1}^{n}\left[\sum_{j=1}^{k}\eta_j T_j(x_i) - A(\eta) + \log h(x_i)\right], \\
&\nabla l(\eta|x) = \sum_{i=1}^{n}T(x_i) - n\nabla A(\eta) \text{ where } T(x_i) = (T_1(x_i), T_2(x_i), \cdots, T_k(x_i)), \\
&\nabla^2 l(\eta|x) = -n\nabla^2 A(\eta).
\end{aligned}$$

Recall Remark 1.4.4 and (1.4.13),

$$A'(\eta) = \mathbb{E}_\eta[T_i(X)],\ A''(\eta) = Var_\eta[T_i(X)] > 0.$$

So $l(\eta|x)$ is strictly concave in η and

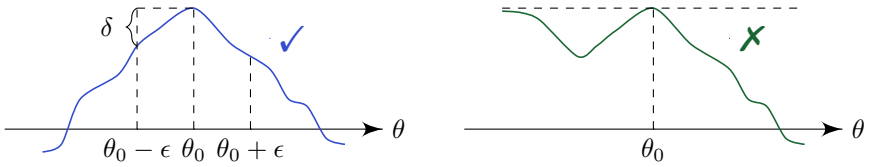
$$0 = \nabla l(\eta|x) \iff \frac{1}{n}\sum_{i=1}^{n} T(x_i) = \nabla A(\eta)$$

has at most one root $\hat{\eta}$. By Theorem 5.2.2, this root exists with probability approaching 1 as $n \to \infty$, and it is consistent for η_0. Since $l(\eta|x)$ is concave, this root $\hat{\eta}$ (when it exists) is the MLE.

Theorem 5.2.3 (General "Consistency Theorem")**.** Suppose

(i) (Separation of maximizer) For any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\mathbb{E}_{\theta_0}[l(\theta_0|X_i)] - \sup_{\theta:\|\theta-\theta_0\|\geq\epsilon} \mathbb{E}_{\theta_0}[l(\theta|X_i)] > \delta. \tag{5.2.7}$$



(ii) (Uniform Law of Large Numbers)

$$\sup_{\theta\in\Omega}\left|\frac{1}{n}\cdot l(\theta|x) - \mathbb{E}_{\theta_0}[l(\theta|X_i)]\right| \xrightarrow{\text{P}} 0 \text{ under } \theta_0. \tag{5.2.8}$$

Then the MLE $\hat{\theta}_n$ is consistent at θ_0.

Proof. Fix any $\epsilon > 0$, let δ be as in condition (i). By condition (ii), with probability approaching 1

$$\sup_{\theta\in\Omega}\left|\frac{1}{n}\cdot l(\theta|x) - \mathbb{E}_{\theta_0}[l(\theta|X_i)]\right| < \frac{\delta}{2}.$$

With this event holds, for any $\|\theta - \theta_0\| \geq \epsilon$,

$$\frac{1}{n}\cdot l(\theta|x) - \frac{\delta}{2} + \delta < \mathbb{E}_{\theta_0}[l(\theta|X_i)] + \delta < \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] < \frac{1}{n}\cdot l(\theta_0|x).$$

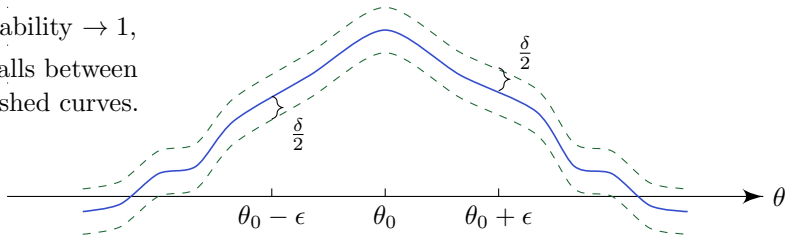
$$\frac{1}{n}\cdot l(\theta|x) < \frac{1}{n}\cdot l(\theta_0|x).$$

So the MLE $\hat{\theta}_n$ satisfies $\|\hat{\theta}_n\theta_0\| < \epsilon$. Hence

$$\mathbb{P}_{\theta_0}[\|\hat{\theta}_n - \theta_0\| < \epsilon] \to 1. \tag{5.2.9}$$

□

With probability $\to 1$, $\frac{1}{n}\cdot l(\theta|x)$ falls between the two dashed curves.



Note 5.2.1. Proofs of condition (ii) typically use a covering net arguement as in the preceding Theorem.

5.3 MLE and Fisher Information

Recap: Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta,\ \theta \in \Omega \subseteq \mathbb{R}^k$. Then the log-likelihood is

$$l(\theta|x) = \sum_{i=1}^{n} l(\theta|x_i),\ l(\theta|x_i) = \log p_\theta(x_i).$$

As $n \to \infty$, under the true parameter θ_0, for each fixed θ,

$$\frac{1}{n} \cdot l(\theta|x) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[l(\theta|X_i)].$$

With certain (mild) regularity assumptions, for any $\epsilon > 0$,

$$\mathbb{P}_{\theta_0}[\text{ there exists a local maximizer } \hat{\theta}_n \text{ of } l(\theta|x) \text{ such that } \|\hat{\theta}_n - \theta_0\| < \epsilon] \to 1.$$

Question 5.3.1. What is the second-order behavior of the MLE $\hat{\theta}_n$ around θ_0 (i.e., its asymptotic distribution around θ_0) ?

Heuristic calculation: For $\theta \in \mathbb{R}$, Taylor expand $l'(\theta|x)$ around θ_0:

$$\begin{aligned} 0 = l'(\hat{\theta}_n|x) &\approx l'(\theta_0|x) + l''(\theta_0|x) \cdot (\hat{\theta}_n - \theta_0) + \cdots \\ \hat{\theta}_n - \theta_0 &\approx -\frac{l'(\theta_0|x)}{l''(\theta_0|x)} \\ \sqrt{n}(\hat{\theta}_n - \theta_0) &\approx \frac{\frac{1}{\sqrt{n}} l'(\theta_0|x)}{-\frac{1}{n} l''(\theta_0|x)}. \end{aligned} \tag{5.3.1}$$

Under the true parameter θ_0:

- For the denominator:

$$-\frac{1}{n} l''(\theta_0|x) = -\frac{1}{n} \sum_{i=1}^{n} l''(\theta_0|x_i) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[-l''(\theta_0|X_i)] \equiv I_1(\theta_0). \tag{5.3.2}$$

- For the numerator:

$$\begin{aligned} \mathbb{E}_{\theta_0} l'(\theta_0|X_i) &= \mathbb{E}_{\theta_0}\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(p_\theta(X_i)) \Big|_{\theta=\theta_0} \right] \\ &= \mathbb{E}_{\theta_0}\left[\frac{\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X_i)\big|_{\theta=\theta_0}}{p_{\theta_0}(X_i)} \right] \\ &= \int \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(X_i)\big|_{\theta=\theta_0}}{p_{\theta_0}(x)} \cdot p_{\theta_0}(x)\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\int p_\theta(x) \right]_{\theta=\theta_0} \text{ (justified in most settings)} \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}[1] = 0. \end{aligned} \tag{5.3.3}$$

  Then by Central Limit Theorem,

$$\begin{aligned} &Z_n \equiv \frac{1}{\sqrt{n}} l'(\theta_0|x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(\theta_0|x_i) \xrightarrow{\text{d}} N(0, I_2(\theta_0)). \\ &\text{where } I_2(\theta_0) = Var_{\theta_0}[l'(\theta_0|X_i)] = \mathbb{E}_{\theta_0}[l'(\theta_0|X_i)^2]. \end{aligned} \tag{5.3.4}$$

– In fact, $I_1(\theta_0) = I_2(\theta_0)$:

$$\begin{aligned}
I_1(\theta_0) &= \mathbb{E}_{\theta_0}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log p_\theta(X_i)\Big|_{\theta=\theta_0}\right] \\
&= \mathbb{E}_{\theta_0}\left[-\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X_i)}{p_\theta(X_i)}\right)\Big|_{\theta=\theta_0}\right] \\
&= \mathbb{E}_{\theta_0}\left[-\frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}p_\theta(X_i)\Big|_{\theta=\theta_0}}{p_{\theta_0}(X_i)} + \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(X_i)\big|_{\theta=\theta_0}\right)^2}{p_\theta(X_i)^2}\right] \\
&= -\mathbb{E}_{\theta_0}\left[\frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}p_\theta(X_i)\Big|_{\theta=\theta_0}}{p_{\theta_0}(X_i)}\right] + \mathbb{E}_{\theta_0}\left[l'(\theta_0|X_i)^2\right] \\
&= \int \frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}p_\theta(x)\Big|_{\theta=\theta_0}}{p_{\theta_0}(x)}\cdot p_{\theta_0}(x)\mathrm{d}x + I_2(\theta_0) \\
&= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left[\int p_\theta(x)\mathrm{d}x\right]_{\theta=\theta_0} + I_2(\theta) \text{ (justfied in most settings)} \\
&= I_2(\theta_0) \equiv I(\theta_0).
\end{aligned} \tag{5.3.5}$$

By Slutsky's Lemma,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} N(0, \frac{I_2(\theta_0)}{I_1(\theta_0)^2}) = N(0, \frac{1}{I(\theta_0)}). \tag{5.3.6}$$

Definition 5.3.1 (Score function and Fisher information)**.** For $\theta_0 \in \mathbb{R}^k$,

$$Z_n \equiv \frac{1}{\sqrt{n}}\nabla l(\theta_0|x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla l(\theta_0|x_i) \in \mathbb{R}^k \tag{5.3.7}$$

is the *score function*.

$$I(\theta_0) = Cov_{\theta_0}[\nabla l(\theta_0|X_i)] \in \mathbb{R}^{k\times k} \tag{5.3.8}$$

is the *Fisher information*.

Note 5.3.1. For $k=1$, $Z_n = \frac{1}{\sqrt{n}}l'(\theta_0|x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n l'(\theta_0|x_i)$, $I(\theta_0) = Var_{\theta_0}[l'(\theta|X_i)]$.

Proposition 5.3.1. Suppose at $\theta = \theta_0$,

$$\nabla_\theta \int p_\theta(x)\mathrm{d}x = \int \nabla_\theta p_\theta(x)\mathrm{d}x, \quad \nabla_\theta^2 \int p_\theta(x)\mathrm{d}x = \int \nabla_\theta^2 p_\theta(x)\mathrm{d}x.$$

Then

- $\mathbb{E}_{\theta_0}[Z_n] = 0$ and $Z_n \xrightarrow{\mathrm{d}} N(0, I(\theta_0))$ under θ_0.
- $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\nabla l(\theta_0|X_i)\cdot\nabla l(\theta_0|X_i)^T\right] = -\mathbb{E}_{\theta_0}[\nabla^2 l(\theta_0|X_i)]$.

Proof. Same as the preceding heuristic sketch. □

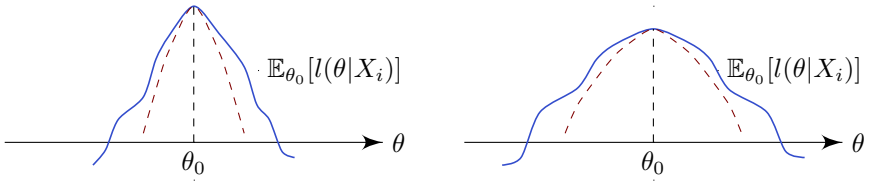
Remark 5.3.1. For exponential family model $p(x|\eta) = \exp(\eta^T T(x) - A(\eta))h(x)$, the conditions $\nabla_\eta \int p(x|\eta)\mathrm{d}x = \int \nabla p(x|\eta)\mathrm{d}x$ and $\nabla^2_\eta \int p_\eta(x)\mathrm{d}x = \int \nabla^2_\eta p_\eta(x)\mathrm{d}x$ hold at $\eta = \eta_0$ in the interior of the natural parameter space Ξ. This may be checked by the Dominated Convergence Theorem.

Interpretation:

(i) $I(\theta_0) = -\nabla^2 \mathbb{E}_{\theta_0}[l(\theta|X_i)]_{\theta=\theta_0}$ is the curvature of the population log-likelihood $\mathbb{E}_{\theta_0}[l(\theta|X_i)]$ at $\theta = \theta_0$. The larger this curvature, the more sharply peaked is the log-likelihood around the true parameter θ_0, so the more "information" X contains about the value of θ. Recall from our heuristic calculation and (5.3.6)

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} N(0, \frac{1}{I(\theta_0)}).$$

The larger the value of $I(\theta_0)$, the smaller the variance of the MLE.



Large $I(\theta_0)$ Small $I(\theta_0)$

(ii) Recall the Kullback–Leibler divergence,

$$D_{KL}(p_{\theta_0}||p_\theta) = \int \log \frac{p_{\theta_0}(x)}{p_\theta(x)} p_{\theta_0}(x)\mathrm{d}x = \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] - \mathbb{E}_{\theta_0}[l(\theta|X_i)].$$

For $\theta \approx \theta_0$, by Taylor expansion,

$$l(\theta|X_i) \approx l(\theta_0|X_i) + \nabla l(\theta_0|X_i)^T(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \nabla^2 l(\theta_0|X_i)(\theta - \theta_0).$$

Take the expectation and we have

$$\begin{aligned}\mathbb{E}_{\theta_0}[l(\theta|X_i)] &\approx \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] + 0 + \frac{1}{2}(\theta - \theta_0)^T \mathbb{E}_{\theta_0}[\nabla^2 l(\theta_0|X_i)](\theta - \theta_0) \\ &= \mathbb{E}_{\theta_0}[l(\theta_0|X_i)] - \frac{1}{2}(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0).\end{aligned}$$

Then

$$D_{KL}(p_{\theta_0}||p_\theta) \approx \frac{1}{2}(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0). \tag{5.3.9}$$

By symmetry

$$\begin{aligned}D_{KL}(p_\theta||p_{\theta_0}) &\approx \frac{1}{2}(\theta_0 - \theta)^T I(\theta)(\theta_0 - \theta) \\ &\approx \frac{1}{2}(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0) \text{ (if } I(\theta) \text{ is continuous in } \theta).\end{aligned} \tag{5.3.10}$$

So for $\theta \approx \theta_0$,

$$D_{KL}(p_{\theta_0}||p_\theta) \approx D_{KL}(p_\theta||p_{\theta_0}) \tag{5.3.11}$$

is approximately a quadratic function $\frac{1}{2}(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0)$ (For $k = 1$, this is $\frac{1}{2}I(\theta_0) \cdot (\theta - \theta_0)^2$).

Example 5.3.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$. $\sigma^2 > 0$ a known value. Here,

$$l(\theta|x_i) = \log p_\theta(x_i) = -\frac{1}{2}\log 2\sigma^2 - \frac{(x_i-\theta)^2}{2\sigma^2}.$$
$$l'(\theta|x_i) = \frac{x_i-\theta}{\sigma^2}.$$
$$l''(\theta|x_i) = -\frac{1}{\sigma^2}.$$

Then the Fisher information is

$$I(\theta) = Var_\theta[l'(\theta|X_i)] = Var_\theta\left[\frac{X_i-\theta}{\sigma^2}\right] = \frac{1}{\sigma^4}\cdot Var_\theta(X_i) = \frac{1}{\sigma^2}.$$
$$I(\theta) = -\mathbb{E}_\theta[l''(\theta|X_i)] = -\mathbb{E}_\theta\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}.$$

The larger the variance σ^2, the less information X contains about θ. Here, the MLE is $\hat{\theta}_n = \bar{X}$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N(0, \sigma^2)$ under θ_0 by Central Limit Theorem. We indeed have $\sigma^2 = \frac{1}{\theta_0}$.

Example 5.3.2. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Poisson(\lambda)$. $p_\lambda(x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$. Here

$$l(\lambda|x_i) = -\log(x_i!) - \lambda + x_i\log\lambda.$$
$$l'(\lambda|x_i) = -1 + \frac{x_i}{\lambda}.$$
$$l''(\lambda|x_i) = -\frac{x_i}{\lambda^2}.$$

Then the Fisher information is

$$I(\lambda) = Var_\lambda[l'(\lambda|X_i)] = \frac{1}{\lambda^2}Var_\lambda[X_i] = \frac{1}{\lambda}.$$
$$I(\lambda) = -\mathbb{E}_\lambda[l''(\lambda|X_i)] = \frac{1}{\lambda^2}\mathbb{E}_\lambda[X_i] = \frac{1}{\lambda}.$$

Here, the MLE is also $\hat{\lambda}_n = \bar{X}$ and $\sqrt{n}(\hat{\lambda}_n - \lambda_0) \xrightarrow{\text{d}} N(0, \lambda_0) = N(0, \frac{1}{I(\lambda_0)})$ under λ_0 by Central Limit Theorem.

Example 5.3.3. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Gamma(a, b)$. Let $\theta = (a, b)$, $p_\theta(x_i) = \frac{1}{\Gamma(a)b^a}x_i^{a-1}e^{-\frac{x_i}{b}}$. Here,

$$l(a, b|x_i) = -\log\Gamma(a) - a\log b + (a-1)\log x_i - \frac{x_i}{b}.$$
$$\nabla l(a, b|x_i) = \left(-\frac{\Gamma'(a)}{\Gamma(a)} - \log b + \log x_i, -\frac{a}{b} + \frac{x_i}{b^2}\right).$$
$$\nabla^2 l(a, b|x_i) = \begin{pmatrix} -\frac{\Gamma''(a)}{\Gamma(a)} + \frac{\Gamma'(a)^2}{\Gamma(a)^2} & -\frac{1}{b} \\ -\frac{1}{b} & \frac{a}{b^2} - \frac{2x_i}{b^3} \end{pmatrix}.$$

Then the Fisher information is

$$I(a, b) = -\mathbb{E}_{a,b}[\nabla^2 l(a, b|X_i)] = \begin{pmatrix} \frac{\Gamma''(a)}{\Gamma(a)} - \frac{\Gamma'(a)^2}{\Gamma(a)^2} & \frac{1}{b} \\ \frac{1}{b} & \frac{a}{b^2} \end{pmatrix}.$$

Here, the MLE solves

$$0 = \frac{1}{n}\nabla l(\theta|x) \implies \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \log x_i = \frac{\Gamma'(a)}{\Gamma(a)} + \log b \\ \frac{1}{n}\sum_{i=1}^{n} x_i = ab \end{cases}$$

There is no closed-form expression for $\hat{\theta}_n = (\hat{a}_n, \hat{b}_n)$. But we expect $\sqrt{n}(\hat{\theta}_n - (a_0, b_0)) \xrightarrow{\text{d}} N(0, I(a_0, b_0)^{-1})$ under (a_0, b_0) by Central Limit Theorem.

5.4 Asymptotic Efficiency of MLE

Theorem 5.4.1 (Cramer-Rao Lower Bound)**.** Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, and let $\delta(X)$ be an unbiased estimator of $g(\theta) \in \mathbb{R}$. Suppose

$$\nabla_\theta \int p_\theta(x)\mathrm{d}x = \int \nabla_\theta p_\theta(x)\mathrm{d}x, \ \nabla_\theta \int \delta(x)p_\theta(x)\mathrm{d}x = \int \delta(x)\cdot\nabla_\theta p_\theta(x)\mathrm{d}x.$$

And $I(\theta)$ is invertible. Then

$$Var_\theta[\delta(X)] \geq \frac{1}{n}\nabla g(\theta)^T \cdot I(\theta)^{-1} \cdot \nabla g(\theta). \tag{5.4.1}$$

Note 5.4.1. For $\theta \in \mathbb{R}$, this gives

$$Var_\theta[\delta(X)] \geq \frac{1}{n}\cdot\frac{g'(\theta)^2}{I(\theta)}. \tag{5.4.2}$$

For $g(\theta) = \theta \in \mathbb{R}$, this further gives

$$Var_\theta[\delta(X)] \geq \frac{1}{nI(\theta)}. \tag{5.4.3}$$

Proof. Let $x = (x_1, x_2, \cdots, x_n)$. Observe that

$$\begin{aligned} \nabla g(\theta) &= \nabla_\theta \int \delta(x)p_\theta(x)\mathrm{d}x \\ &= \int \delta(x)\cdot\nabla_\theta p_\theta(x)\mathrm{d}x \\ &= \int \delta(x)\cdot\nabla_\theta l(\theta|x)\cdot p_\theta(x)\mathrm{d}x. \end{aligned}$$

Then, for any vector $a \in \mathbb{R}^k$,

$$\begin{aligned} \nabla g(\theta)^T a &= \int \delta(x)(\nabla l(\theta|x)^T a)p_\theta(x)\mathrm{d}x \\ &= Cov_\theta[\delta(X), \nabla l(\theta|X)^T a] \quad (\text{Since } \mathbb{E}_\theta[\nabla_\theta l(\theta|X)] = 0). \\ &\leq \sqrt{Var_\theta[\delta(X)]}\cdot\sqrt{Var_\theta\left[\nabla_\theta l(\theta|X)^T a\right]} \quad (\text{By Cauchy–Schwarz Inequality.}) \end{aligned}$$

So

$$Var_\theta[\delta(X)] \geq \frac{(g(\theta)^T a)^2}{Var_\theta\left[\nabla_\theta l(\theta|X)^T a\right]}.$$

When $x = (x_1, x_2, \cdots, x_n)$ are i.i.d., we have

$$\begin{aligned} Var_\theta \left[\nabla_\theta l(\theta|X)^T a\right] &= Var_\theta \left[\sum_{i=1}^n \nabla_\theta l(\theta|X)^T a\right] \\ &= n \cdot Var_\theta \left[\nabla_\theta l(\theta|X_i)^T a\right] \\ &= n \cdot a^T Cov_\theta[\nabla_\theta l(\theta|X_i)]a \\ &= n \cdot a^T I(\theta) a. \end{aligned}$$

Therefore,

$$Var_\theta[\delta(X)] \geq \frac{(g(\theta)^T a)^2}{n \cdot a^T I(\theta) a}.$$

Let $a = I(\theta)^{-1} \cdot \nabla g(\theta)$. Then

$$Var_\theta[\delta(X)] \geq \frac{(\nabla g(\theta)^T a)^2}{n \cdot a^T I(\theta) a} = \frac{1}{n} \cdot \nabla g(\theta)^T I(\theta)^{-1} \cdot \nabla g(\theta). \tag{5.4.4}$$

□

Interpretation: This is an information theoretic lower bound for the variance achievable by *any* unbiased estimaotr. And for the MLE $\hat{\theta}_n$ of θ, as $n \to \infty$,

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{d}} N\left(0, I(\theta)^{-1}\right). \\ &\hat{\theta}_n \approx N\left(\theta, \frac{1}{n} \cdot I(\theta)^{-1}\right). \end{aligned}$$

So $\hat{\theta}_n$ asymptotically achieves the Cramer-Rao Lower Bound. Furthermore, for the plug-in estimator $g(\hat{\theta}_n)$, by Delta Method,

$$\begin{aligned} &\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{\text{d}} N\left(0, \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)\right). \\ &g(\hat{\theta}_n) \approx N\left(g(\theta), \frac{1}{n} \cdot \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)\right). \end{aligned}$$

Again, $g(\hat{\theta}_n)$ asymptotically achieves the Cramer-Rao Lower Bound.

Later in this class, we'll describe in an asymptotic sense that "$\mathbb{E}_\theta[n \cdot (\delta(X) - g(\theta))^2]$ is at least as large as $\nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)$ for *any* estimator $\delta(X)$", including possibly biased estimators.

Definition 5.4.1 (Asymptotically efficient)**.** An estimator $\hat{\theta}_n$ of θ is *asymptotically efficient* if, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{d}} N(0, I(\theta)^{-1}). \tag{5.4.5}$$

An estimator $\delta(X)$ of $g(\theta)$ is asymptotically efficient if

$$\sqrt{n}(\delta(X) - g(\theta)) \xrightarrow{\text{d}} N\left(0, \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)\right). \tag{5.4.6}$$

Theorem 5.4.2 (Asymptotic Normality and Efficiency of MLE)**.** Consider the model $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \Omega \subseteq \mathbb{R}^k$. Let θ_0 be the true parameter. Suppose that

(i) Ω contains an open neighborhood ω of θ_0, and $l(\theta|x_i)$ is three-times differentiable in $\theta \in \omega$.

(ii) $\nabla_\theta \int p_\theta(x)\mathrm{d}x = \int \nabla_\theta p_\theta(x)\mathrm{d}x$ and $\nabla_\theta^2 \int p_\theta = \int \nabla_\theta^2 p_\theta(x)\mathrm{d}x$.

(iii) $I(\theta)$ is invertible and continuous at θ_0.

(iv) For some function $m(x_i)$ and any $j_1, j_2, j_3 \in \{1, 2, \cdots, k\}$

$$\left|\frac{\partial^3}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}} l(\theta|x_i)\right| \leq m(x_i) \text{ for all } \theta \in \omega, \text{ and } \mathbb{E}_{\theta_0}[m(X_i)] \leq \infty.$$

Let $\hat{\theta}_n$ be any root of $0 = \nabla l(\theta|x)$ which is asymptotically consistent for θ_0, and let $Z_n = \frac{1}{\sqrt{n}}\nabla l(\theta|x)$ be the score function. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1} Z_n + r_n \text{ where } r_n \xrightarrow{\text{P}} 0 \text{ under } \theta_0. \tag{5.4.7}$$

Furthermore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N(0, I(\theta_0)^{-1}). \tag{5.4.8}$$

Proof. Let $\Psi_j(\theta|x) = \frac{\partial}{\partial\theta_j} l(\theta|x)$, so $\nabla l(\theta|x) = (\Psi_1(\theta|x), \Psi_2(\theta|x), \cdots, \Psi_k(\theta|x))$. Let $\Psi_j(\theta|x_i) = \frac{\partial}{\partial\theta_j} l(\theta|x_i)$, so $\Psi_j(\theta|x) \sum_{i=1}^n \Psi_j(\theta|x_i)$ .Then $0 = \Psi_j(\hat{\theta}_n|x)$ for each $j = 1, 2, \cdots, k$. With probability approaching 1,

$$0 = \Psi_j(\hat{\theta}_n|x) = \Psi_j(\theta_0|x) + \nabla\Psi_j(\theta_0|x)^T \cdot (\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T \cdot \nabla^2\Psi_j(\tilde{\theta}|x) \cdot (\hat{\theta}_n - \theta_0)$$

where $\tilde{\theta}$ is on the line segment between $\hat{\theta}_n$ and θ_0.

Then

$$\begin{aligned}\frac{1}{\sqrt{n}}\Psi_j(\theta_0|x) &= \left(-\frac{1}{n}\nabla\Psi_j(\theta_0|x)^T - \frac{1}{2n}(\hat{\theta}_n - \theta_0)^T\nabla^2\Psi_j(\tilde{\theta}|x)\right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) \\ &\equiv \left(-\frac{1}{n}\nabla\Psi_j(\theta_0|x)^T - R_j(x)^T\right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0)\end{aligned} \tag{5.4.9}$$

For $R_j(x)$,

- For each entry of $\frac{1}{n}\nabla^2\Psi_j(\tilde{\theta}|x) = \frac{1}{n}\sum_{i=1}^n \nabla^2\Psi_j(\tilde{\theta}|x_i)$ ($k \times k$ entries in total) is bounded by $\bar{m}(x) = \frac{1}{n}\sum_{i=1}^n m(x_i)$ for each $j = 1, 2, \cdots, k$.
- For any matrix $M \in \mathbb{R}^{k\times k}$ and vector $v \in \mathbb{R}^k$, if $|M_{ij}| \leq \bar{m}$ for all entries i, j. Let $M_i =$ row i of M. Then

$$\|Mv\|^2 = \sum_{i=1}^k (M_i^T v)^2 \leq \sum_{i=1}^k \|M_i^T\|^2 \cdot \|v\|^2 \leq \sum_{i=1}^k k\bar{m}^2 \cdot \|v\|^2 = k^2\bar{m}^2 \cdot \|v\|^2.$$

So

$$\|Mv\| \leq k\bar{m} \cdot \|v\| \implies \|R_j(x)\| \leq \frac{1}{2}k\bar{m}(x) \cdot \|\hat{\theta}_n - \theta_0\|. \tag{5.4.10}$$

For any $\epsilon > 0$, $T > 0$,

$$\begin{aligned}
\mathbb{P}_{\theta_0}\left[\|R_j(X)\| > \epsilon\right] &\leq \mathbb{P}_{\theta_0}\left[\frac{1}{2}k\bar{m}(x) \cdot \|\hat{\theta}_n - \theta_0\| > \epsilon\right] \\
&\leq \mathbb{P}_{\theta_0}\left[\frac{1}{2}k\bar{m}(x) > T\right] + \mathbb{P}_{\theta_0}\left[\|\hat{\theta}_n - \theta_0\| > \frac{\epsilon}{T}\right] \\
&\leq \frac{k}{2T}\mathbb{E}_{\theta_0}\left[\bar{m}(X)\right] + \mathbb{P}_{\theta_0}\left[\|\hat{\theta}_n - \theta_0\| > \frac{\epsilon}{T}\right] \text{ by Markov's Inequality}
\end{aligned}$$

Since $\hat{\theta}_n$ is consistent, $\mathbb{P}_{\theta_0}\left[\|\hat{\theta}_n - \theta_0\| > \frac{\epsilon}{T}\right] \to 0$.

$$\lim_{n\to\infty} \sup \mathbb{P}_{\theta_0}\left[\|R_j(X)\| > \epsilon\right] \leq \frac{k}{2T}\mathbb{E}_{\theta_0}\left[\bar{m}(X)\right]. \tag{5.4.11}$$

As $\mathbb{E}_{\theta_0} < \infty$ and T is arbitary, we have $\mathbb{P}_{\theta_0}\left[\|R_j(X)\| > \epsilon\right] \to 0$ (i.e., $R_j(X) \xrightarrow{\text{P}} 0$.)

Putting this together for $j = 1, 2, \cdots, k$

$$\frac{1}{\sqrt{n}}\nabla l(\theta_0|x) = \left(-\frac{1}{n}\nabla^2 l(\theta_0|x) - R(x)\right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) \text{ where } R(X) = \begin{pmatrix} R_1(X)^T \\ R_2(X)^T \\ \cdots \\ R_k(X)^T \end{pmatrix}.$$

We know that

- $R(X) \xrightarrow{\text{P}} 0$ under θ_0.
- $-\frac{1}{n}\nabla^2 l(\theta_0|x) \xrightarrow{\text{P}} \mathbb{E}_{\theta_0}[-\nabla^2 l(\theta_0|x)] = I(\theta_0)$.
- $I(\theta)$ is invertible and continuous at θ_0.

Then

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &= \left(-\frac{1}{n}\nabla^2 l(\theta_0|x) - R(x)\right)^{-1} \cdot Z_n \\
&= (I(\theta_0)^{-1} + \tilde{r}_n) \cdot Z_n \text{ where } \tilde{r}_n \xrightarrow{\text{P}} 0.
\end{aligned} \tag{5.4.12}$$

Since $Z_n \xrightarrow{\text{d}} N(0, I(\theta_0))$ by Central Limit Theorem.

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1} \cdot Z_n + r_n \text{ where } r_n \xrightarrow{\text{P}} 0 \\
&\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N(0, I(\theta_0)^{-1}) \text{ (by Slutsky's Lemma).}
\end{aligned} \tag{5.4.13}$$

□

Remark 5.4.1. Three times differentiability of $l(\theta|x)$ leads to a straightforward proof, but this condition is not necessary: See *van der Vaart* Theorem 5.13 for weaker conditions.

Corollary 5.4.1. If the MLE $\hat{\theta}_n$ is the unique root of $0 = \nabla l(\theta|x)$ with probability approaching 1, then it is asymptotically efficient.

Question 5.4.1. What if $0 = \nabla l(\theta|x)$ has multiple roots?

Answer 5.4.1. In practice, we can apply Newton's Method:

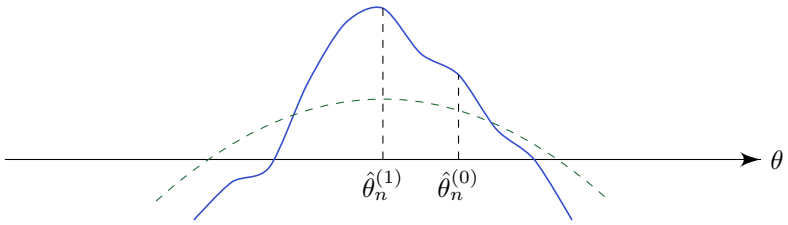
- Find an initial estimator $\hat{\theta}_n^{(0)}$.
- Perform Newton's Method: Given estimate $\hat{\theta}^{(k)}$, approximate $l(\theta|x)$ by

$$l\left(\hat{\theta}_n^{(k)}|x\right)+\nabla l\left(\hat{\theta}_n^{(k)}|x\right)\left(\theta-\hat{\theta}_n^{(k)}\right)+\frac{1}{2}\left(\theta-\hat{\theta}_n^{(k)}\right)^T\cdot\nabla^2 l(\hat{\theta}_n^{(k)}|x)\cdot\left(\theta-\hat{\theta}_n^{(k)}\right). \tag{5.4.14}$$

  Maximize this approximate over θ and get

$$\hat{\theta}_n^{(k+1)} = \hat{\theta}_n^{(k)} - \left[\nabla^2 l(\hat{\theta}_n^{(k)}|x)\right]^{-1}\cdot\left[\nabla l(\hat{\theta}_n^{(k)}|x)\right]. \tag{5.4.15}$$

- Iterate until convergence.



Theorem 5.4.3. Under the conditions of Theorem 5.4.2, suppose $\tilde{\theta}_n$ is an initial estimate of θ such that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converges in distribution. Then

$$\hat{\theta}_n = \tilde{\theta}_n - \left[\nabla^2 l(\tilde{\theta}_n|x)\right]^{-1}\left[\nabla l(\tilde{\theta}_n|x)\right] \tag{5.4.16}$$

obtained from *one* Newton step is asymptotically efficient, i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N(0, I(\theta_0)^{-1}) \text{ under } \theta_0. \tag{5.4.17}$$

Proof sketch. We can rewrite (5.4.16):

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}(\tilde{\theta}_n - \theta_0) - \left[\frac{1}{n}\nabla^2 l(\tilde{\theta}_n|x)\right]^{-1}\left[\frac{1}{\sqrt{n}}\nabla l(\tilde{\theta}_n|x)\right]$$

- For $\frac{1}{\sqrt{n}}\nabla l(\tilde{\theta}_n|x)$,

$$\begin{aligned}\frac{1}{\sqrt{n}}\nabla l(\tilde{\theta}_n|x) &= \frac{1}{\sqrt{n}}\nabla l(\theta_0|x) + \frac{1}{\sqrt{n}}\nabla^2 l(\theta_0|x)(\tilde{\theta}_n - \theta_0) + \cdots \\ &= Z_n - I(\theta_0)\cdot\sqrt{n}(\tilde{\theta}_n - \theta_0) + r_n \text{ where } r_n \xrightarrow{\text{P}} 0.\end{aligned}$$

- For $\frac{1}{n}\nabla^2 l(\tilde{\theta}_n|x)$,

$$\begin{aligned}\frac{1}{n}\nabla^2 l(\tilde{\theta}_n|x) &= \frac{1}{n}\nabla^2 l(\theta_0|x) + \cdots \\ &= -I(\theta_0) + r_n \text{ where } r_n \xrightarrow{\text{P}} 0.\end{aligned}$$

Then

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= \sqrt{n}(\tilde{\theta} - \theta_0) - (-I(\theta_0)^{-1} + r_n) \cdot (Z_n - I(\theta_0) \cdot \sqrt{n}(\tilde{\theta}_n - \theta_0) + r_n) \\ &= I(\theta_0)^{-1} Z_n + r_n \xrightarrow{\text{d}} N(0, I(\theta_0)^{-1}).\end{aligned} \tag{5.4.18}$$

□

Question 5.4.2. How to get $\sqrt{n}$-consistent initialization $\tilde{\theta}_n$?

Answer 5.4.2. One way is method of moments.

Proposition 5.4.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \mathbb{R}^k$. Pick k statistics $T_1, T_2, \cdots, T_k$ and set $\mu_1 = \mathbb{E}_\theta[T_1(X_i)], \mu_2 = \mathbb{E}_\theta[T_2(X_i)], \cdots, \mu_k = \mathbb{E}_\theta[T_k(X_i)]$. If $\theta = g(\mu)$ for a differentiable function $g : \mathbb{R}^k \to \mathbb{R}^k$, then the so-called *method of moments* estimate $\hat{\theta} = g(\hat{\mu})$ where $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n T_1(X_i), \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n T_2(X_i), \cdots, \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n T_k(X_i)$ is asymptotically normal.

Proof. By Central Limit Theorem,

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{\text{d}} N(0, \Sigma) \text{ for some } \Sigma \in \mathbb{R}^{k \times k}.$$

Then by Delta Method,

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{\text{d}} N(0, D^T \Sigma D) \text{ for } D = Dg(\mu).$$

□

Example 5.4.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} Gamma(a, b)$. Then

$$\left\{ \begin{array}{l} u_1 \equiv \mathbb{E}[X_i] = ab \\ u_2 \equiv \mathbb{E}\left[X_i^2\right] = ab^2 + a^2 b^2 \end{array} \right\} \implies b = \frac{\mu_2 - \mu_1^2}{\mu_1},\ a = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

Let $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n X_i,\ \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$, then by Proposition 5.4.1., $\hat{a} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}, \hat{b} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1}$ is asymptotically normal.

6 Local Asymptotics

6.1 Hypothesis Testing and Contiguity

Example 6.1.1. Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$, $\sigma^2 > 0$ known. Test

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1 \ (\theta_1 > \theta_0).$$

The likelihood ratio is

$$L(x) = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta_1)^2}{2\sigma^2}}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta_0)^2}{2\sigma^2}}} = \exp\left(\frac{n(\theta_1-\theta_0)}{\sigma^2} \cdot \bar{x} - \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2}\right)$$

The most powerful test rejects for large $\bar{X}$.

- Under θ_0, $\bar{X} \sim N(\theta_0, \frac{\sigma^2}{n})$.
- Under θ_1, $\bar{X} \sim N(\theta_1, \frac{\sigma^2}{n})$.

Suppose θ_0, θ_1 are fixed as $n \to \infty$, and ϕ_n rejects H_0 when $\bar{X} > \frac{\theta_0+\theta_1}{2}$. Then

$$\mathbb{P}_{\theta_0}\left[\bar{X} > \frac{\theta_0+\theta_1}{2}\right] = \mathbb{P}_{\theta_0}\left[\sqrt{n}(\bar{X}-\theta_0) > \frac{\sqrt{n}}{2}(\theta_1-\theta_0)\right] \to 0.$$
$$\mathbb{P}_{\theta_1}\left[\bar{X} > \frac{\theta_0+\theta_1}{2}\right] = \mathbb{P}_{\theta_1}\left[\sqrt{n}(\bar{X}-\theta_0) > \frac{\sqrt{n}}{2}(\theta_0-\theta_1)\right] \to 1.$$

This test is asymptotically perfect: We can tell whether the truth is H_0 or H_1 with probability approaching 1, as $n \to \infty$.

Interpretation: The hypothesized distributions

- $P_n : \ X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta_0, \sigma^2)$
- $Q_n : \ X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta_1, \sigma^2)$

for the data $X_1, X_2, \cdots X_n \in \mathbb{R}^n$ are "asymptotically do not overlap" if there is a n-dependent set $E_n \subset \mathbb{R}^n$ (namely the rejection region of ϕ_n such as $\{(x_1, x_2, \cdots, x_n) : \bar{x} > \frac{\theta_0+\theta_1}{2}\}$) such that

$$P_n\left[(X_1, X_2, \cdots, X_n) \in E_n\right] \to 0, \ Q_n\left[(X_1, X_2, \cdots, X_n) \in E_n\right] \to 1.$$

Suppose now θ_0 is fixed, but $\theta_1 \equiv \theta_{1,n}$ changes with n. If $\theta_{1,n} - \theta_0 \gg \frac{1}{\sqrt{n}}$ (i.e., $\sqrt{n}(\theta_{1,n}-\theta_0) \to \infty$), the above still holds. The test is no longer asymptotically perfect if $\theta_{1,n}$ is such that $\sqrt{n}(\theta_{1,n}-\theta_0)$ remains bounded.

Definition 6.1.1 (Local alternative and local parameter)**.** For fixed $\theta_0 \subseteq \mathbb{R}^k$, $\theta_{1,n} \equiv \theta_0 + \frac{h}{\sqrt{n}}$ for a fixed vector $h \in \mathbb{R}^k$ (as $n \to \infty$) is called a *local alternative* to θ_0. h is the *local parameter*.

Perspective: For fixed θ_0 in the interior of the parameter space $\Omega \subseteq \mathbb{R}^k$, write $\theta = \theta_0 + \frac{h}{\sqrt{n}}$ and think about $h = \sqrt{n}(\theta - \theta_0)$ as an n-dependent local reparametrization of the model. We'll consider some asymptotics as $n \to \infty$, not just at θ_0 (i.e., $h = 0$), but for general $h \in \mathbb{R}^k$.

Motivations:

- To test $\theta \in \Omega \subseteq \mathbb{R}^k$ in a parametric model,

$$H_0 : \theta = \theta_0 \text{ v.s. } H_1 : \theta = \theta_{1,n}$$

  if $\|\theta_{1,n} - \theta_0\| \gg \frac{1}{\sqrt{n}}$, then the most "resonable" tests can distinguish H_0 from H_1 asymptotically perfectly. Differences in power between different test are exhibited at local alternatives $\theta_{1,n} - \theta_0 \asymp \frac{1}{\sqrt{n}}$.

- To compare different estimators of θ, it is sometimes deceptive to compare their asymptotic behavior at each $\theta \in \Omega$. We will want to understand their behavior uniformly over "small" regions of Ω, of size $\frac{1}{\sqrt{n}}$. We will parametrize this region by $\theta_0 + \frac{h}{\sqrt{n}}$, where θ_0 is a fixed point at the center of this region.

Definition 6.1.2 (Contiguous and mutually contiguous)**.** Let $\mathcal{X}_n$ be a sample space, and let P_n and Q_n be two probability distributions on $\mathcal{X}_n$. As $n \to \infty$, Q_n is *contiguous* to P_n if, for any $E_n \subset \mathcal{X}_n$ such that $P_n(E_n) \to 0$, we have $Q_n(E_n) \to 0$. We write $Q_n \triangleleft P_n$. Q_n and P_n are *mutually contiguous* if $Q_n \triangleleft P_n$ and $P_n \triangleleft Q_n$. We write $Q_n \lhd\rhd P_n$.

Testing interpretation: Let E_n be the rejection region of some test ϕ_n, for

$$H_0 : X \sim P_n \text{ v.s. } H_1 : X \sim Q_n$$

Then $Q_n \triangleleft P_n$ means if type I error $\to 0$ under P_n as $n \to \infty$, then the power under $Q_n \to 0$ as $n \to \infty$ as well.

Proposition 6.1.1. $Q_n \triangleleft P_n$ if and only if, for any statistics $T_n : \mathcal{X}_n \to \mathbb{R}$ where $T_n \xrightarrow{\text{P}} 0$ under P_n, also $T_n \xrightarrow{\text{P}} 0$ under Q_n.

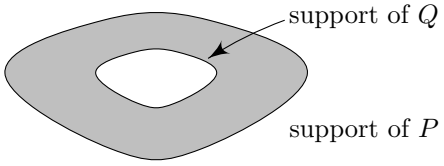
Proof.

(i) $\Longrightarrow$: Suppose $Q_n \triangleleft P_n$. Fix $\epsilon > 0$ and let $E_n = \{x : |T_n(x)| > \epsilon\}$. If $T_n \xrightarrow{\text{P}} 0$ under P_n, then $P_n(E_n) \to 0$. So $Q_n(E_n) \to 0$. This holds for any $\epsilon > 0$, so $T_n \xrightarrow{\text{P}} 0$ under Q_n.

(ii) $\Longleftarrow$: Suppose $T_n \xrightarrow{\text{P}} 0$ under P_n implies $T_n \xrightarrow{\text{P}} 0$ under Q_n. For any $E_n \subset \mathcal{X}_n$, let $T_n(x) = \mathbb{1}\{x \in E_n\}$. If $P_n(E_n) \to 0$, then $T_n \xrightarrow{\text{P}} 0$ under P_n. So $T_n \xrightarrow{\text{P}} 0$ under Q_n, meaning $Q_n(E_n) \to 0$.

□

Definition 6.1.3 (Absolutely continuous)**.** For fixed distributions P, Q on $\mathcal{X}$, Q is *absolutely continuous* with respect to P if, for any $E \subset \mathcal{X}$ where $P(E) = 0$, we have $Q(E) = 0$.



Proposition 6.1.2. Let P, Q have densities $p(x), q(x)$ with respect to a measure μ on $\mathcal{X}$. Then the following are equivalent:

(i) Q is absolutely continuous with respect to P.

(ii) $Q\left(\left\{x : \frac{p(x)}{q(x)} = 0\right\}\right) = 0$.

(iii) $\mathbb{E}_P\left[\frac{q(x)}{p(x)}\right] = 1$.

Proof.

(i) It is equivalent to: For μ-a.e. $x \in \mathcal{X}$ where $p(x) = 0$, we must also have $q(x) = 0$.

(ii) We rewrite the left hand side,

$$Q\left(\left\{x : \frac{p(x)}{q(x)}\right\}\right) = \int \mathbb{1}\{p(x) = 0\} \cdot q(x) \mathrm{d}\mu(x).$$

This is 0 if and only if (i) holds.

(iii) We rewrite the left hand side,

$$\begin{aligned}
\mathbb{E}_P\left[\frac{q(x)}{p(x)}\right] &= \int \frac{q(x)}{p(x)} \cdot \mathbb{1}\{p(x) > 0\} \cdot p(x) \mathrm{d}\mu(x) \\
&= \int \mathbb{1}\{p(x) > 0\} \cdot q(x) \mathrm{d}\mu(x) \\
&= 1 - \int \mathbb{1}\{p(x) = 0\} \cdot q(x) \mathrm{d}\mu(x) \\
&= 1 - Q\left(\left\{x : \frac{p(x)}{q(x)}\right\}\right).
\end{aligned}$$

This is 1 if and only if (ii) holds.

□

For contiguity $Q_n \triangleleft P_n$, there is an "asymptotic version" of the preceding result.

Theorem 6.1.1 (Le Cam's First Lemma)**.** Let P_n, Q_n be (n-indexed sequences of) probability distributions on $\mathcal{X}_n$, having densities p_n and q_n with respect to some common measure μ_n. Suppose that,

$$\text{Under } Q_n : \frac{p_n(x)}{q_n(x)} \xrightarrow{\mathrm{d}} U.$$

$$\text{Under } P_n : \frac{q_n(x)}{p_n(x)} \xrightarrow{\mathrm{d}} V.$$

Then the following are equivalent

(i) $Q_n \triangleleft P_n$.

(ii) $\mathbb{P}[U = 0] = 0$.

(iii) $\mathbb{E}[V] = 1$.

Partial proof. We'll show (ii) $\implies$ (i) and also (iii) $\implies$ (i).
In both cases, suppose $E_n \subset \mathcal{X}_n$ satisfies $P_n(E_n) \equiv \alpha_n \to 0$. Write

$$\begin{aligned} Q_n(E_n) &= Q_n\left(E_n \bigcap \{x : q_n(x) \le t p_n(x)\}\right) + Q_n\left(E_n \bigcap \{x : q_n(x) > t p_n(x)\}\right) \\ &\equiv \mathrm{I} + \mathrm{II} \quad \text{for a constant } t > 0. \end{aligned}$$

To bond I,

$$\begin{aligned} \mathrm{I} &= \int \mathbb{1}\{x \in E_n\} \cdot \mathbb{1}\{q_n(t) \le t p_n(x)\} \cdot q_n(x) \mathrm{d}\mu_n(x) \\ &\le \int \mathbb{1}\{x \in E_n\} \cdot t p_n(x) \mathrm{d}\mu_n(x) \\ &= t \cdot P_n(E_n) \\ &= t \cdot \alpha_n \to 0 \quad \text{for any fixed } t > 0 \text{ as } n \to \infty. \end{aligned}$$

It remains to bound II.

- To prove (ii) $\implies$ (i):

  $$\begin{aligned} \mathrm{II} &\le Q_n\left(\{q_n(x) > t p_n(x)\}\right) \\ &= Q_n\left(\{x : \frac{p_n(x)}{q_n(x)} < \frac{1}{t}\}\right) \quad \text{(We may assume } q_n(x) > 0 \text{ when } X \sim Q_n\text{).} \end{aligned}$$

  Let U_n be a random variable with the distribution of $\frac{p_n(x)}{q_n(x)}$ when $X \sim Q_n$.
  The condition (ii) says $U_n \xrightarrow{\mathrm{d}} U$ where $\mathbb{P}[U = 0] = 0$. Fix any $\epsilon > 0$. Pick t sufficiently large such that

  - $\mathbb{P}\left[U < \frac{1}{t} < \epsilon\right]$.
  - $\frac{1}{t}$ is a continuity point of the CDF of U.

  Then

  $$\mathbb{P}\left[U_n < \frac{1}{t}\right] \to \mathbb{P}\left[U < \frac{1}{t}\right] < \epsilon.$$

  So

  $$\sup \mathrm{II} \le \epsilon \text{ where } \epsilon \text{ is arbitary.} \implies \mathrm{II} \to 0.$$

  This shows $Q_n(E_n) \to 0$ (i.e., $Q_n \vartriangle P_n$).

- To prove (iii) $\implies$ (i):

  $$\begin{aligned} \mathrm{II} &\le Q_n\{x : q_n(x) > t p_n(x)\}) \\ &= 1 - Q_n(\{q_n(x) \le t p_n(x)\}) \\ &= 1 - \int \mathbb{1}\{q_n(x) \le t p_n(x)\} \cdot q_n(x) \mathrm{d}\mu_n(x) \\ &= 1 - \int_{x : p_n(x) > 0} \mathbb{1}\left\{\frac{q_n(x)}{p_n(x)} \le t\right\} \cdot \frac{q_n(x)}{p_n(x)} \cdot p_n(x) \mathrm{d}\mu_n(x). \end{aligned}$$

  Let V_n be a random variable with the distribution of $\frac{q_n(x)}{p_n(x)}$ where $X \sim P_n$.
  Then the above is

  $$\mathrm{II} \le 1 - \mathbb{E}\left[\mathbb{1}\{V_n \le t\} \cdot V_n\right].$$

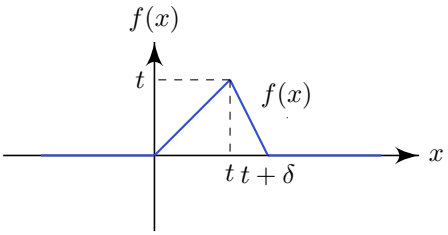
  The condition (iii) says $V_n \xrightarrow{\mathrm{d}} V$ where $\mathbb{E}[V] = 0$. Fix any $\epsilon > 0$. Pick t sufficiently large such that

- $\mathbb{E}\left[\mathbb{1}\{V \le t\} \cdot V\right] > 1 - \epsilon$.
- t is a continuity point of the CDF of V.

Claim 6.1.1. $\mathbb{E}\left[\mathbb{1}\{V_n \le t\} \cdot V_n\right] \to \mathbb{E}\left[\mathbb{1}\{V \le t\} \cdot V\right] > 1 - \epsilon$

Proof. Let $f(x)$ be a continuous approximation of $\mathbb{1}\{x \le t\} \cdot x$:

$$f(x) = \begin{cases} 0 & x \le 0 \\ x & 0 < x \le t \\ \in (0,t) & t < x \le t + \delta \\ 0 & x > t + \delta \end{cases} \quad \text{for some } \delta > 0.$$



Since f is continuous and bounded and $V_n \xrightarrow{\text{d}} V$, we have

$$\mathbb{E}[f(V_n)] \to \mathbb{E}[f(V)].$$

On the other hand,

$$\begin{aligned} \mathbb{E}[f(V_n)] &\ge \mathbb{E}\left[\mathbb{1}\{V_n \le t\} \cdot V_n\right] \\ &\ge \mathbb{E}\left[f(V_n)\right] - t \cdot \mathbb{P}\left[V_n \in (t, t+\delta)\right] \quad \text{(Since } V_n \ge 0\text{)}. \end{aligned}$$

Take δ such that $t + \delta$ is also a continuity point of the CDF of V. Then

$$\mathbb{P}[V_n \in (t, t+\delta)] \to \mathbb{P}[V \in (t, t+\delta)]$$

Then

$$\begin{aligned} \mathbb{E}[f(V)] &\ge \lim_{n\to\infty} \sup \mathbb{E}\left[\mathbb{1}\{V_n \le t\} \cdot V_n\right] \\ &\ge \lim_{n\to\infty} \inf \mathbb{E}\left[\mathbb{1}\{V_n \le t\} \cdot V_n\right] \\ &\ge \mathbb{E}\left[f(V_n)\right] - t \cdot \mathbb{P}\left[V_n \in (t, t+\delta)\right]. \end{aligned}$$

Since δ is arbitrarily small, this yields

$$\mathbb{E}\left[\mathbb{1}\{V_n \le t\} \cdot V_n\right] \to \mathbb{E}\left[\mathbb{1}\{V \le t\} \cdot V\right] > 1 - \epsilon.$$

□

By Claim 6.1.1,

$$\sup \text{II} \le 1 - (1 - \epsilon) = \epsilon \text{ where } \epsilon \text{ is arbitary.} \implies \text{II} \to 0.$$

This shows $Q_n(E_n) \to 0$ (i.e., $Q \triangle P$).

To prove (i) $\implies$ (ii) and (i) $\implies$ (iii), see *Theory of Point Estimation* Theorem 12.3.2, *van der Vaart* Lemma 6.4. □

Note 6.1.1. If $\frac{q_n(x)}{p_n(x)} \xrightarrow{\text{d}} V$ under P_n, $\mathbb{P}[V=0]=0$, $\mathbb{E}[V]=1$, then $Q_n \lhd\rhd P_n$.

Remark 6.1.1. Actually, we don't need to assume $\frac{p_n(x)}{q_n(x)} \xrightarrow{\text{d}} U$ under Q_n and $\frac{q_n(x)}{p_n(x)} \xleftrightarrow{\text{d}} V$ under P_n. If $\frac{p_n(x)}{q_n(x)} \xrightarrow{\text{d}} U$ along any subsequence, or $\frac{q_n(x)}{p_n(x)} \xrightarrow{\text{d}} V$ along any subsequence, then the preceding still shows $Q_n \triangle P_n$. See *van der Vaart* Theorem 6.4 for a full statement.

Example 6.1.2. Consider $\mathcal{X}_n = \mathbb{R}^n$.

- $P_n =$ joint law of $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} n(\theta_0, \sigma^2)$.
- $Q_n =$ joint law of $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta_1, \sigma^2)$ where $\theta_1 \equiv \theta_0 + \frac{h}{\sqrt{n}}$ for a fixed local parameter $h > 0$.

We are going to show $Q_n \lhd\rhd P_n$.

$$\begin{aligned}\frac{q_n(x)}{p_n(x)} &= \exp\left(\frac{n(\theta_1-\theta_0)}{\sigma^2}\cdot\bar{x} - \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2}\right)\\ &= \exp\left(\frac{h}{\sigma^2}\cdot\sqrt{n}\bar{x} - \frac{h}{\sigma^2}\cdot\sqrt{n}\theta_0 - \frac{h^2}{2\sigma^2}\right)\\ &= \exp\left(\frac{h}{\sigma^2}\cdot\sqrt{n}(\bar{x}-\theta_0) - \frac{h^2}{2\sigma^2}\right).\end{aligned}$$

Under P_n, we have

$$\frac{q_n(x)}{p_n(x)} \overset{\text{L}}{=} e^W \quad \text{where } W \sim N(-\frac{h^2}{2\sigma^2}, \frac{h^2}{\sigma^2}).$$

The mean of W is $-\frac{1}{2}$ times the variance.
Note: For any $Z \sim N(\mu, \sigma^2)$, $\mathbb{E}\left[e^Z\right] = e^{\mu+\frac{1}{2}\sigma^2}$. So

$\mathbb{E}\left[e^W\right] = 1 \implies$ By Le Cam's First Lemma ((iii) $\implies$ (i)), $Q_n \triangle P_n$.
$\mathbb{P}\left[e^W = 0\right] = 0 \implies$ By Le Cam's First Lemma ((ii) $\implies$ (i)), $P_n \triangle Q_n$.

So $P_n \lhd\rhd Q_n$.

Corollary 6.1.1. Suppose P_n, Q_n are such that $\log\frac{q_n(x)}{p_n(x)} \to W$ under P_n, where $W \sim N(\mu, \sigma^2)$. Then $P_n \triangle Q_n$ always, and $Q_n \triangle P_n$ if and only if $\mu = -\frac{1}{2}\sigma^2$.

6.2 Local Asymptotic Normality

Setup:. Consider $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \Omega \subseteq \mathbb{R}^k$. Let θ_0 be in the interior of Ω. Our goal is to show under mild conditions, $P_n : X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$ and $Q_n : X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta_0+\frac{h}{\sqrt{n}}}$ are mutually contiguous.

Heuristic calculation: Write $l(\theta|x) = \sum_{i=1}^{n} l(\theta|x_i) = \sum_{i=1}^{n} \log p_\theta(x_i)$. Applying a Taylor expansion, we expect for fixed $h \in \mathbb{R}^k$

$$\begin{aligned}\log L_{n,h} \equiv \log \frac{\prod_{i=1}^{n} p_{\theta_0 + \frac{h}{\sqrt{n}}}(x_i)}{\prod_{i=1}^{n} p_{\theta_0}(x_i)} &= l(\theta_0 + \frac{h}{\sqrt{n}}|x) - l(\theta_0|x) \\ &\approx \frac{h}{\sqrt{n}}^T \cdot \nabla l(\theta_0|x) + \frac{1}{2}\frac{h}{\sqrt{n}}^T \cdot \nabla^2 l(\theta_0|x) \cdot \frac{h}{\sqrt{n}} \\ &= \left[\frac{1}{\sqrt{n}}\nabla l(\theta_0|x)\right]^T h - \frac{1}{2}h^T \left[-\frac{1}{n}\nabla^2 l(\theta_0|\ x)\right] h.\end{aligned}$$

Under θ_0:

- $\frac{1}{\sqrt{n}}\nabla l(\theta_0|x) \xrightarrow{\text{d}} N(0, I(\theta_0))$.
- $-\frac{1}{n}\nabla^2 l(\theta_0|x) \xrightarrow{\text{P}} I(\theta_0)$.

So we expect

$$\log L_{n,h} \xrightarrow{\text{d}} N\left(-\frac{1}{2}h^T I(\theta_0)h, h^T I(\theta_0)h\right).$$

$$L_{n,h} \xrightarrow{\text{d}} e^W \text{ where } W \sim N\left(-\frac{1}{2}h^T I(\theta_0)h, h^T I(\theta_0)h\right)$$

Then by Le Cam's First Lemma,

$$\left.\begin{array}{ll} \circ & \mathbb{P}\left[e^w = 0\right] = 0 \\ \circ & \mathbb{E}\left[e^W\right] = 1 \end{array}\right\} \implies Q_n \lhd\rhd P_n.$$

To make this rigorous:
The following weak regularity condition is all that's needed.

Definition 6.2.1 (Quadratic mean differentiable)**.** Let $\{P_\theta : \theta \in \Omega\}$ have densties p_θ with respect to a common measure μ. Let $\theta_0 \in \Omega \subseteq \mathbb{R}^k$ be an interior point. This model is *quadratic mean differentiable* (q.m.d.) at θ_0 if there is a function $\dot{l} : \mathcal{X} \to \mathbb{R}^k$ such that

$$\frac{1}{\|t\|^2}\int_{\mathcal{X}} \left(\sqrt{p_{\theta_0+t}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2}\sqrt{p_{\theta_0}(x)}\dot{l}(x)^T t\right)^2 \mathrm{d}\mu(x) \to 0 \text{ as } t \to 0. \tag{6.2.1}$$

Interpretation: Suppose $\theta \to \sqrt{p_\theta(x)}$ is differentiable. By Taylor expansion,

$$\sqrt{p_{\theta+t}(x)} = \sqrt{p_\theta(x)} + \frac{1}{2\sqrt{p_\theta(x)}}\nabla p_\theta(x)^T t + O(t^2).$$

We should identify

$$\frac{1}{2\sqrt{p_\theta(x)}}\nabla p_\theta(x)^T t = \frac{1}{2}\sqrt{p_{\theta_0}(x)}\dot{l}(x)^T t.$$

Then

$$\dot{l}(x) = \frac{\nabla p_\theta(x)}{p_\theta(x)} = \nabla \log p_\theta(x).$$

In this context, $\dot{l}(x)$ is the simple-sample score. The usual score is $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}(x_i)$ and the Fisher information is $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\dot{l}(X_i)\dot{l}(X_i)^T\right]$.

Proposition 6.2.1 (Sufficient condition for q.m.d.)**.** If $\theta \to \sqrt{p_\theta(x)}$ is continuously differentiable in θ for μ-a.e. x, and the entries of $I(\theta) = \mathbb{E}\left[\nabla l(\theta|X_i)\nabla l(\theta|X_i)^T\right]$ are well-defined and continuous in θ, then the model is q.m.d. with

$$\frac{1}{2}\sqrt{p_{\theta_0}(x)}\dot{l}(x) = \nabla\sqrt{p_\theta(x)}\big|_{\theta=\theta_0}.$$

So

$$\dot{l}(x) = \nabla \log p_\theta(x)\big|_{\theta=\theta_0} = \frac{\nabla p_\theta(x)\big|_{\theta=\theta_0}}{p_{\theta_0}(x)}.$$

Proof. See *van der Vaart* Lemma 7.6. □

Note 6.2.1. q.m.d. holds for

- Exponential families.
- Location models $p_\theta(x) = f(x-\theta)$ where $\theta \in \mathbb{R}$, $f : \mathbb{R}^k \to \mathbb{R}$ is a fixed function such that $\sqrt{f}$ is absolutely continuous and $I \equiv \int \frac{f'(x)^2}{f(x)}\mathrm{d}x < \infty$. In particular: $Cauchy(\theta, 1), Laplace(\theta, 1)$.

Theorem 6.2.1 (Local Asymptotic Normality)**.** Suppose $\{P_\theta : \theta \in \Omega\}$ is q.m.d. at $\theta_0 \in \Omega \subseteq \mathbb{R}^k$. Define the score $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}(x_i)$, and the Fisher information $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\dot{l}(X_i)\dot{l}(X_i)^T\right]$. Then for each fixed $h \in \mathbb{R}^k$, as $n \to \infty$,

(i) $\log L_{n,h} \equiv \log \frac{\prod_{i=1}^{n} p_{\theta_0+\frac{h}{\sqrt{n}}}(x_i)}{\prod_{i=1}^{n} p_{\theta_0}(x_i)} = Z_n^T h - \frac{1}{2}h^T I(\theta_0)h + r_n$, where $r_n \xrightarrow{\mathrm{P}} 0$ under θ_0.

(ii) Therefore, under θ_0, $L_{n,h} \xrightarrow{\mathrm{d}} e^W$, $W \sim N(-\frac{1}{2}h^T I(\theta_0)h, h^T I(\theta_0)h)$.

Corollary 6.2.1. By Le Cam's First Lemma, $P^n_{\theta_{\theta_0+\frac{h}{\sqrt{n}}}} \lhd\rhd P^n_{\theta_0}$.

Proof. Fix $h \in \mathbb{R}^k$. Write as shorthand $p(x) \equiv p_{\theta_0}(x), p_n(x) \equiv p_{\theta_0+\frac{h}{\sqrt{n}}}(x)$.

– Step 0: $\mathbb{E}_{\theta_0}\left[Z_n^T h\right] = 0$, $h^T I(\theta_0)h < \infty$.

Proof. Write $\langle f, g\rangle_\mu = \int f(x)g(x)\mathrm{d}\mu(x)$, $\|f\|_\mu^2 = \langle f, f\rangle_\mu = \int f(x)^2\mathrm{d}\mu(x)$. Apply q.m.d. condition with $t = \frac{h}{\sqrt{n}}$,

$$\frac{n}{\|h\|^2}\left\|\sqrt{p_n} - \sqrt{p} - \frac{1}{2}\sqrt{p}\; \dot{l}^T\frac{h}{\sqrt{n}}\right\|_\mu^2 \to 0.$$

So

$$\left\|\sqrt{n}\,(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2}\sqrt{p}\; \dot{l}^T\frac{h}{\sqrt{n}}\right\|_\mu^2 \to 0.$$

- To prove $h^T I(\theta_0)h < \infty$,

$$\begin{aligned} h^T I(\theta_0)h &= \mathbb{E}_{\theta_0}\left[h^T \dot{l}(X_i)\dot{l}(X_i)^T h\right] \\ &= \mathbb{E}_{\theta_0}\left[\left(\dot{l}(X_i)^T h\right)^2\right] \\ &= \int \left(\dot{l}(x_i)^T h\right)^2 p(x_i)\mathrm{d}\mu(x) \\ &= \left\|\sqrt{p}\ \dot{l}^T h\right\|_\mu^2. \end{aligned}$$

Since

$$\begin{aligned} \left\|\sqrt{p}\ \dot{l}^T h\right\|_\mu &\le 2\cdot\left\|\sqrt{n}\left(\sqrt{p_n}-\sqrt{p}\right)\right\|_\mu + 2\cdot\left\|\frac{1}{2}\sqrt{p}\ \dot{l}^T h - \sqrt{n}\left(\sqrt{p_n}-\sqrt{p}\right)\right\| \\ &\le 2\cdot\|\sqrt{n}\sqrt{p_n}\|_\mu + 2\cdot\|\sqrt{n}\sqrt{p}\|_\mu + 2\cdot\left\|\frac{1}{2}\sqrt{p}\ \dot{l}^T h - \sqrt{n}\left(\sqrt{p_n}-\sqrt{p}\right)\right\| \\ &\equiv 2\cdot\|\sqrt{n}\sqrt{p_n}\|_\mu + 2\cdot\|\sqrt{n}\sqrt{p}\|_\mu + 2\epsilon_n \text{ where } \epsilon_n \to 0 \text{ by q.m.d.} \\ &= 2\cdot\left(\int n\cdot p_n\ \mathrm{d}\mu\right)^{\frac{1}{2}} + 2\cdot\left(\int n\cdot p\ \mathrm{d}\mu\right)^{\frac{1}{2}} + 2\epsilon_n \\ &= 4\cdot\sqrt{n} + 2\epsilon_n. \end{aligned}$$

Since $\sqrt{p}\ \dot{l}^T h$ is independent of n, this implies $\|\sqrt{p}\ \dot{l}^T h\|_\mu < \infty$. Then $h^T I(\theta_0)h < \infty$.

- To prove $\mathbb{E}_{\theta_0}\left[l(X_i)^T h\right]$, we first consider

$$\mathbb{E}_{\theta_0}\left[\dot{l}(X_i)^T h\right] = \int \left(\dot{l}(x_i)p(x)\right)\mathrm{d}\mu(x) = \left\langle \frac{1}{2}\sqrt{p}\ \dot{l}^T h, 2\sqrt{p}\right\rangle_\mu.$$

Note:

$$\begin{aligned} &\|2\sqrt{p} - \left(\sqrt{p}+\sqrt{p_n}\right)\|_\mu = \|\sqrt{p}-\sqrt{p_n}\|_\mu \\ &= \frac{1}{\sqrt{n}}\left\|\frac{1}{2}\sqrt{p}\ \dot{l}^T h\right\|_\mu + \frac{1}{\sqrt{n}}\left\|\sqrt{n}\left(\sqrt{p_n}-\sqrt{p}\right) - \frac{1}{2}\sqrt{p}\ \dot{l}^T h\right\|_\mu \to 0. \end{aligned}$$

By q.m.d. and $\sqrt{p}\ \dot{l}^T h$ is independent of n, $\|2\sqrt{p} - \left(\sqrt{p}+\sqrt{p_n}\right)\|_\mu \to 0$. Then

$$\begin{aligned} \mathbb{E}_{\theta_0}\left[\dot{l}(X_i)^T h\right] &= \left\langle \frac{1}{2}\sqrt{p}\ \dot{l}^T h, 2\sqrt{p}\right\rangle_\mu \\ &= \lim_{n\to\infty}\left\langle \sqrt{n}(\sqrt{p_n}-\sqrt{p}), \sqrt{p}+\sqrt{p_n}\right\rangle_\mu \\ &= \int \sqrt{n}(\sqrt{p_n}-\sqrt{p})(\sqrt{p}+\sqrt{p_n})\mathrm{d}\mu \\ &= \int \sqrt{n}(p_n - p)\mathrm{d}\mu = 0. \end{aligned}$$

So this shows $\mathbb{E}_{\theta_0}\left[\dot{l}(X_i)^T h\right] = 0$, also $\mathbb{E}_{\theta_0}\left[Z_n^T h\right] = 0$.

□

– Step 1: Split $\log L_{n,h} \equiv \log \frac{\prod_{i=1}^{n} p_{\theta_0+\frac{h}{\sqrt{n}}}(x_i)}{\prod_{i=1}^{n} p_{\theta_0}(x_i)} = 2\sum_{i=1}^{n} \log \sqrt{\frac{p_n(x_i)}{p(x_i)}}$.

Let $W_n(x_i) = 2\left(\sqrt{\frac{p_n(x_i)}{p(x_i)}} - 1\right)$, then apply Taylor expansion

$$\log(1+x) = x - \frac{x^2}{2} + x^2 R(x) \text{ where } R(x) \to 0 \text{ as } x \to 0.$$

Then

$$\begin{aligned}
\log L_{n,h} &= 2 \cdot \sum_{i=1}^{n} \log\left(1 + \frac{1}{2}W_n(x_i)\right) \\
&= 2 \cdot \left(\sum_{i=1}^{n}\left(\frac{1}{2}W_n(x_i) - \frac{1}{8}W_n(x_i)^2 + \frac{1}{4}W_n(x_i)^2 \cdot R\left(\frac{1}{2}W_n(x_i)\right)\right)\right) \\
&= \sum_{i=1}^{n} W_n(x_i) - \frac{1}{4}\cdot\sum_{i=1}^{n} W_n(x_i)^2 + \frac{1}{2}\cdot\sum_{i=1}^{n} W_n(x_i)^2 \cdot R\left(\frac{1}{2}W_n(x_i)\right).
\end{aligned}$$

– Step 2: $\sum_{i=1}^{n} W_n(x_i) = Z_n^T h - \frac{1}{4}\cdot h^T I(\theta_0) h + r_n$ where $r_n \xrightarrow{\text{P}} 0$ under θ_0.

Proof. Define $r_n = \sum_{i=1}^{n} W_n(x_i) - Z_n^T h + \frac{1}{4}\cdot h^T I(\theta_0) h$.

∘ To prove $\mathbb{E}_{\theta_0}[r_n] \to 0$,

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\sum_{i=1}^{n} W_n(X_i)\right] &= n \cdot \mathbb{E}_{\theta_0}\left[W_n(X_i)\right] \\
&= n \cdot \int 2\left(\sqrt{\frac{p_n(x_i)}{p(x_i)}} - 1\right) \cdot p \, \mathrm{d}\mu \\
&= n\left(\int 2\sqrt{p_n \cdot p} \, \mathrm{d}\mu - 2\right) \\
&= n \cdot \left(\int 2\sqrt{p_n \cdot p} \, \mathrm{d}\mu - \int p_n \, \mathrm{d}\mu - \int p \, \mathrm{d}\mu\right) \\
&= -n \cdot \left(\int (\sqrt{p_n} - \sqrt{p})^2 \, \mathrm{d}\mu\right) \\
&= -\|\sqrt{n}(\sqrt{p_n} - \sqrt{p})\|_\mu^2 \\
&\to -\left\|\frac{1}{2}\sqrt{p}\; l^T h\right\|_\mu^2 \\
&= -\frac{1}{4}\int \left(l^T h\right)^2 \cdot p \, \mathrm{d}\mu \\
&= -\frac{1}{4}\mathbb{E}_{\theta_0}\left[\left(l^T h\right)^2\right] = -\frac{1}{4}h^T I(\theta_0) h.
\end{aligned}$$

Since $E_{\theta_0}\left[Z_n^T h\right] = 0$, we have $\mathbb{E}_{\theta_0}[r_n] \to 0$.

- To prove $Var_{\theta_0}(r_n) \to 0$.

$$\begin{aligned}
Var_{\theta_0}[r_n] &= Var_{\theta_0}\left[\sum_{i=1}^n W_n(x_i) - Z_n^T h\right] \\
&= Var_{\theta_0}\left[\sum_{i=1}^n \left(W_n(X_i) - \frac{1}{\sqrt{n}}\dot{l}(X_i)^T h\right)\right] \\
&= n \cdot Var_{\theta_0}\left[\left(W_n(X_i) - \frac{1}{\sqrt{n}}\dot{l}(X_i)^T h\right)\right] \\
&\leq n\mathbb{E}_{\theta_0}\left[\left(W_n(X_i) - \frac{1}{\sqrt{n}}\dot{l}(X_i)^T h\right)^2\right] \\
&= n\dot{\int}\left(2\left(\frac{p_n}{p}\right) - \frac{1}{\sqrt{n}}\ \dot{l}^T h\right)^2 \cdot p\ \mathrm{d}\mu. \\
&= 4 \cdot \int \left(\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2}\sqrt{p}\ \dot{l}^T h\right)^2 \mathrm{d}\mu \\
&= 4 \cdot \left\|\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2}\sqrt{p}\ \dot{l}^T h\right\|_\mu^2 \to 0.
\end{aligned}$$

So $r_n \xrightarrow{\mathrm{P}} 0$. □

– Step 3: $\sum_{i=1}^n W_n(x_i)^2 \xrightarrow{\mathrm{P}} h^T I(\theta_0) h$ under θ_0.

Proof. Define $A_n(x_i) = W_n(x_i) - \frac{1}{\sqrt{n}}\ \dot{l}(x_i)^T h$. By the variance computation of Step 2, we have $n \cdot \mathbb{E}_{\theta_0}\left[A_n(x_i)^2\right] \to 0$ under θ_0 $(*)$. Write

$$\begin{aligned}
W_n(x_i)^2 &= \left(A_n(x_i) + \frac{1}{\sqrt{n}}\ \dot{l}(x_i)^T h\right)^2 \\
&= \frac{1}{n}\left(\dot{l}(x_i)^T h\right)^2 + \frac{2}{\sqrt{n}}\dot{l}(x_i)^T h \cdot A_n(x_i) + A_n(x_i)^2.
\end{aligned}$$

Then

$$\sum_{i=1}^n W_n(x_i)^2 = \sum_{i=1}^n \frac{1}{n}\left(\dot{l}(x_i)^T h\right)^2 + \sum_{i=1}^n A_n(x_i)^2 + 2 \cdot \sum_{i=1}^n \frac{1}{\sqrt{n}}\dot{l}(x_i)^T h \cdot A_n(x_i).$$

Since

- $\sum_{i=1}^n \frac{1}{n}\left(\dot{l}(x_i)^T h\right)^2 \xrightarrow{\mathrm{P}} h^T I(\theta_0) h$ by Weak Law of Large Numbers.
- $\sum_{i=1}^n A_n(x_i)^2 \xrightarrow{\mathrm{P}} 0$ by $(*)$ and Markov's Inequality.
- $\sum_{i=1}^n \frac{1}{\sqrt{n}}\dot{l}(x_i)^T h \cdot A_n(x_i) \xrightarrow{\mathrm{P}} 0$ by Cauchy–Schwarz Inequality.

We conclude that $\sum_{i=1}^n W_n(x_i)^2 \xrightarrow{\mathrm{P}} h^T I(\theta_0) h$. □

– Step 4: $\sum_{i=1}^{n} W_n(x_i)^2 \cdot R\left(\frac{1}{2}W_n(x_i)\right) \xrightarrow{\text{P}} 0$.

Proof.

$$\sum_{i=1}^{n} W_n(x_i)^2 \cdot R\left(\frac{1}{2}W_n(x_i)\right) \leq \sum_{i=1}^{n} W_n(x_i)^2 \cdot \max_{i=1}^{n}\left|R\left(\frac{1}{2}W_n(x_i)\right)\right|$$

By Step 3, $\sum_{i=1}^{n} W_n(x_i)^2 \xrightarrow{\text{P}} h^T I(\theta_0)h$. So by continuous mapping, it suffices to show $\max_{i=1}^{n}|W_n(x_i)| \xrightarrow{\text{P}} 0$ to show $R\left(\frac{1}{2}W_n(x_i)\right) \xrightarrow{\text{P}} 0$.

If we apply

$$W_n(x_i)^2 = \left(\frac{1}{\sqrt{n}}\,\dot{l}(x_i)^T h + A_n(x_i)\right)^2 \leq \frac{2}{n}\cdot\left(\dot{l}(x_i)^T h\right)^2 + 2\cdot A_n(x_i)^2.$$

Then

$$\begin{aligned}\mathbb{P}_{\theta_0}\left[\max_{i=1}^{n}|W_n(x_i)| \geq \epsilon\right] &\leq n\cdot\mathbb{P}_{\theta_0}\left[W_n(x_i)^2 \geq \epsilon^2\right] \\ &\leq n\left(\mathbb{P}_{\theta_0}\left[\frac{1}{n}\cdot\left(\dot{l}(x_i)^T h\right)^2 \geq \frac{\epsilon^2}{4}\right] + \mathbb{P}_{\theta_0}\left[A_n(x_i)^2 \geq \frac{\epsilon^2}{4}\right]\right).\end{aligned}$$

Note: For any random variable $Y \geq 0$, let $Z = Y\cdot\mathbb{1}\{Y \geq t\}$,

$$\mathbb{P}[Y \geq t] = \mathbb{P}[Z \geq t] \leq \frac{1}{t}\cdot\mathbb{E}[Z] = \frac{1}{t}\cdot\mathbb{E}\left[Y\cdot\mathbb{1}\{Y \geq t\}\right]. \tag{6.2.2}$$

Then

- $n\cdot\mathbb{P}_{\theta_0}\left[A(x_i)^2 \geq \frac{\epsilon^2}{4}\right] \leq \frac{4n}{\epsilon^2}\cdot\mathbb{E}_{\theta_0}\left[A(x_i)^2\right] \to 0$ by Markov's Inequality.
- $n\cdot\mathbb{P}_{\theta_0}\left[\frac{1}{n}\cdot\left(\dot{l}(x_i)^T h\right)^2 \geq \frac{\epsilon^2}{4}\right] \leq n\cdot\frac{4}{n\epsilon^2}\mathbb{E}_{\theta_0}\left[\left(\dot{l}(x_i)^T h\right)^2\cdot\mathbb{1}\left\{\dot{l}(x_i)^T h \geq \frac{n\epsilon^2}{4}\right\}^2\right]$ by (6.2.2). Since $\mathbb{E}_{\theta_0}\left[\left(\dot{l}(x_i)^T h\right)^2\right] = h^T I(\theta_0)h < \infty$, this $\to 0$ as $n \to \infty$ by Dominated Convergence Theorem.

This shows $\max_{i=1}^{n}|W_n(x_i)| \xrightarrow{\text{P}} 0$ and then $\sum_{i=1}^{n} W_n(x_i)^2\cdot R\left(\frac{1}{2}W_n(x_i)\right) \xrightarrow{\text{P}} 0$. □

– Combining Step 1 - 4:

$$\begin{aligned}\log L_{n,h} &= Z_n^T h - \frac{1}{4}h^T I(\theta_0)h - \frac{1}{4}h^T I(\theta_0)h + r_n \\ &= Z_n^T h - \frac{1}{2}h^T I(\theta_0)h + r_n \text{ where } r_n \xrightarrow{\text{P}} 0 \text{ under } \theta_0\end{aligned} \tag{6.2.3}$$

This shows (i) in Theorem 6.2.1. Under θ_0,

$$\begin{aligned}&Z_n^T h \xrightarrow{\text{d}} N\left(0, h^T I(\theta_0)h\right) \quad \text{by Central Limit Theorem.} \\ &\log L_{n,h} \xrightarrow{\text{d}} N\left(-\frac{1}{2}h^T I(\theta_0)h, h^T I(\theta_0)h\right).\end{aligned} \tag{6.2.4}$$

This shows (ii) in Theorem 6.2.1.

□

6.3 Local Alternatives, Aymptotic Power

Example 6.3.1 (Neyman-Pearon test)**.** Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \mathbb{R}^k$. Assume the model is q.m.d. at θ_0. Test

$$H_0 : \theta = \theta_0 \text{ v.s. } H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}} \text{ for fixed } h \in \mathbb{R}^k.$$

By Neyman-Pearson Lemma, the most powerful test rejects H_0 when

$$\log L_{n,h} \equiv \log \prod_{i=1}^{n} \frac{p_{\theta_0 + \frac{h}{\sqrt{n}}}(x_i)}{p_{\theta_0}(x_i)} > l_{n,h}^{(1-\alpha)}$$

where $l_{n,h}^{(1-\alpha)}$ is the $(1-\alpha)$-quantile of the distribution of $\log L_{n,h}$ under H_0. By Theorem 6.2.1., under $H_0 : \theta = \theta_0$

$$\log L_{n,h} \xrightarrow{\text{d}} N\left(-\frac{1}{2}h^T I(\theta_0)h, h^T I(\theta_0)h\right) \equiv N\left(-\frac{1}{2}\sigma_h^2, \sigma_h^2\right).$$

Then $l_{n,h}^{(1-\alpha)}$ converges to the $(1-\alpha)$-quantile of this normal limit,

$$l_{n,h}^{(1-\alpha)} \to -\frac{1}{2}\sigma_h^2 + \sigma_h \cdot z^{(1-\alpha)}.$$

Question 6.3.1. What is the asymptotic power of this test?

Answer 6.3.1. To compute the asymptotic power, we need to understand the distribution of $L_{n,h}$ not under H_0, but under H_1.

Heuristic calculation: Let $p_n(l)$ be the density of $\log L_{n,h}(x)$ under H_0, and $p(l)$ be the density of its limit $N\left(-\frac{1}{2}\sigma_h^2, \sigma_h^2\right)$. Assuming $p_{\theta_0 + \frac{h}{\sqrt{n}}}$ and p_{θ_0} have common support, we have for any function $f : \mathcal{X}^n \to \mathbb{R}$

$$\begin{aligned}
\mathbb{E}_{\theta_0 + \frac{h}{\sqrt{n}}}\left[f(X_1, X_2, \cdots, X_n)\right] &= \int f(x_1, x_2, \cdots, x_n) \cdot \prod_{i=1}^{n} p_{\theta_0 + \frac{h}{\sqrt{n}}}(x_i) \mathrm{d}x_i \\
&= \int f(x_1, x_2, \cdots, x_n) \cdot \frac{\prod_{i=1}^{n} p_{\theta_0 + \frac{h}{\sqrt{n}}}(x_i)}{\prod_{i=1}^{n} p_{\theta_0}(x_i)} \cdot \prod_{i=1}^{n} p_{\theta_0}(x_i) \mathrm{d}x_i \\
&= \mathbb{E}_{\theta_0}\left[f(X_1, X_2, \cdots, X_n) \cdot L_{n,h}(x)\right].
\end{aligned}$$

Specialize this to $f(x_1, x_2, \cdots, x_n) = g\left(\log L_{n,h}\right)$ for any $g : \mathbb{R} \to \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{E}_{\theta_0 + \frac{h}{\sqrt{n}}}\left[g\left(\log L_{n,h}\right)\right] &= \mathbb{E}_{\theta_0}\left[f(X_1, X_2, \cdots, X_n) \cdot e^{\log L_{n,h}(x)}\right] \\
&= \int_{\mathbb{R}} g(l) \cdot e^l \cdot p_n(l) \mathrm{d}l \to \int_{\mathbb{R}} g(l) \cdot e^l \cdot p(l) \mathrm{d}l \quad \text{as } n \to \infty.
\end{aligned}$$

This holds for every (nice enough) $g : \mathbb{R} \to \mathbb{R}$, so the limit distribution of $\log L_{n,h}$ under $H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}$ has density $e^l \cdot p(l)$. We have

$$p(l) = \frac{1}{\sqrt{2\pi\sigma_h^2}} \cdot e^{-\frac{\left(l + \frac{1}{2}\sigma_h^2\right)^2}{2\sigma_h^2}} = \frac{1}{\sqrt{2\pi\sigma_h^2}} \cdot e^{-\frac{l^2 + l\sigma_h^2 + \frac{\sigma_h^4}{4}}{2\sigma_h^2}}.$$

$$e^l p(l) = \frac{1}{\sqrt{2\pi\sigma_h^2}} \cdot e^{-\frac{l^2 - l\sigma_h^2 + \frac{\sigma_h^4}{4}}{2\sigma_h^2}} = \frac{1}{\sqrt{2\pi\sigma_h^2}} \cdot e^{-\frac{\left(l - \frac{1}{2}\sigma_h^2\right)^2}{2\sigma_h^2}}.$$

So we expect under $H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}$,

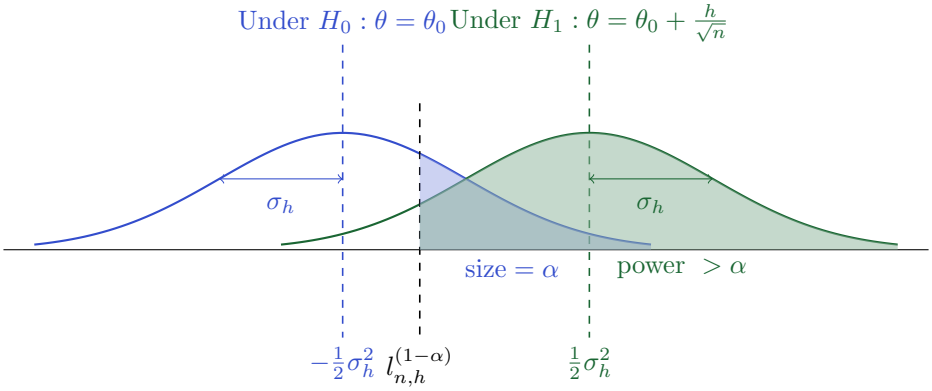
$$\log L_{n,h} \xrightarrow{\text{d}} N\left(\frac{1}{2}\sigma_h^2, \sigma_h^2\right).$$

Then as $n \to \infty$, the asymptotic power is

$$\begin{aligned}
\mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[\log L_{n,h} > l_{n,h}^{(1-\alpha)}\right] &\to \mathbb{P}\left[N\left(\frac{1}{2}\sigma_h^2, \sigma_h^2\right) > -\frac{1}{2}\sigma_h^2 + \sigma_h \cdot z^{(1-\alpha)}\right] \\
&= \mathbb{P}\left[\frac{1}{2}\sigma_h^2 + \sigma_h \cdot Z > -\frac{1}{2}\sigma_h^2 + \sigma_h \cdot z^{(1-\alpha)}\right] \quad \text{where } Z \sim N(0,1) \\
&= \mathbb{P}\left[Z > -\sigma_h + z^{(1-\alpha)}\right] \\
&= \tilde{\Phi}\left(-\sigma_h + z^{(1-\alpha)}\right) \quad \text{where } \tilde{\Phi}(x) = \mathbb{P}[Z > x] \text{ for } Z \sim N(0,1) \\
&\equiv \tilde{\Phi}\left(-\frac{\text{mean shift}}{\text{standardized deviation}} + z^{(1-\alpha)}\right)
\end{aligned}$$

This power is larger than α as long as $\sigma_h^2 = h^T I(\theta_0) h$.

Limit Distributions of $\log L_{n,h}$



To formalize this argument:

Lemma 6.3.1. Random variables Y_n converge in distribution to Y if and only if, for any non-negative continuous function f, we have

$$\lim_{n\to\infty} \inf \mathbb{E}\left[f(Y_n)\right] \geq \mathbb{E}\left[f(Y)\right].$$

Proof. See *van der Vaart* Lemma 2.2. □

Theorem 6.3.1 (Le Cam's Third Lemma)**.** Suppose P_n, Q_n are probability distributions on $\mathcal{X}_n$ such that $Q_n \triangleleft P_n$. Let $T_n : \mathcal{X}_n \to \mathbb{R}$ be any statistic such that under P_n,

$$(T_n, \log L_n) \xrightarrow{\text{d}} (T, W) \text{ where } L_n(x) \equiv \frac{q_n(x)}{p_n(x)} \text{ is the likelihood ratio.}$$

Then under Q_n,

$$(T_n, \log L_n) \xrightarrow{\text{d}} (\tilde{T}, \tilde{W}) \text{ where } \mathbb{P}[(\tilde{T}, \tilde{W}) \in B] = \mathbb{E}\left[\mathbb{1}\{(T, W) \in B\} \cdot e^W\right].$$

Note 6.3.1.

- If (T, W) has joint density $p(t, w)$, then $(\tilde{T}, \tilde{W})$ has joint density $p(t, w) \cdot e^w$.
- If $T_n = \log L_n$, then this reduces to

$$\log L_n \xrightarrow{\text{d}} \tilde{W} \text{ where } \mathbb{P}[\tilde{W} \in B] = \mathbb{E}\left[\mathbb{1}\{W \in B\} \cdot e^W\right].$$

If W has density $p(w)$, then $\tilde{T} = \tilde{W}$ has density $p(w) \cdot e^w$.

Corollary 6.3.1. Suppose under P_n,

$$\begin{pmatrix} T_n \\ \log L_n \end{pmatrix} \xrightarrow{\text{d}} N\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma_2^2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}\right) \equiv N\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma_2^2 \end{pmatrix}, \Sigma\right).$$

Then under Q_n,

$$T_n \xrightarrow{\text{d}} N\left(\mu + \sigma_{1,2}, \sigma_1^2\right).$$

Proof of Corollary 6.3.1. Note that marginally,

$$\log L_n \to N\left(-\frac{1}{2}\sigma_2^2, \sigma_2^2\right).$$

So by Le Cam's First Lemma, $Q_n \triangleleft P_n$. Under P_n, $(T_n, \log L_n)$ converges in dsitribution to the density

$$p(t, w) = \frac{1}{2\pi\left|\Sigma\right|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{1}{2}\left(t - \mu, w + \frac{1}{2}\sigma_2^2\right)^T \Sigma^{-1}\left(t - \mu, w + \frac{1}{2}\sigma_2^2\right)\right\}.$$

Then by Le Cam's Third Lemma, $(T_n, \log L_n)$ under Q_n has limit density $p(t, w) \cdot e^w$. A direct computation show that under Q_n,

$$\begin{pmatrix} T_n \\ \log L_n \end{pmatrix} \xrightarrow{\text{d}} N\left(\begin{pmatrix} \mu + \sigma_{1,2} \\ \frac{1}{2}\sigma_2^2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}\right).$$

So in particularly marginally for T_n, the limit is $N\left(\mu + \sigma_{1,2}, \sigma_1^2\right)$.

□

Proof of Theorem 6.3.1. Let's take $B = \mathbb{R}^2$.

(i) Since $Q_n \triangleleft P_n$,

$$\mathbb{P}\left[(\tilde{T}, \tilde{W}) \in \mathbb{R}^2\right] = \mathbb{E}\left[e^W\right] = 1 \text{ by Le Cam's First Lemma.}$$

Then the probability distribution of $(\tilde{T}, \tilde{W})$ is well-defined.

(ii) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be any non-negative continuous function, then

$$\begin{aligned} \mathbb{E}_{Q_n}\left[f(T_n, \log L_n)\right] &\geq \mathbb{E}_{Q_n}\left[\mathbb{1}\{p_n > 0\} \cdot f(T_n, \log L_n)\right] \\ &= \mathbb{E}_{P_n}\left[f(T_n, \log L_n) \cdot e^{\log L_n}\right] \end{aligned}$$

Note: $g(t, w) = f(t, w) \cdot e^w$ is also a non-negative continuous function. So

$$\begin{aligned} \lim_{n\to\infty} \inf \mathbb{E}_{P_n}\left[f(T_n, \log L_n) \cdot L_n\right] &= \lim_{n\to\infty} \inf \mathbb{E}_{P_n}\left[g(T_n, \log L_n)\right] \\ &\geq \mathbb{E}[g(T, W)] \text{ by Lemma 6.3.1} \\ &= \mathbb{E}\left[f(T, W) \cdot e^W\right] = \mathbb{E}\left[f(\tilde{T}, \tilde{W})\right]. \end{aligned}$$

This holds for any non-negative continuous function f, so by Lemma 6.3.1, under Q_n,

$$(T_n, \log L_n) \xrightarrow{\text{d}} (\tilde{T}, \tilde{W}).$$

□

Remark 6.3.1. Typically, we'll apply Theorem 6.3.1 by noting

– Under $P_{\theta_0}^n$, the likelihood ratio $L_{n,h} = \prod_{i=1}^n \frac{p_{\theta_0+\frac{h}{\sqrt{n}}}(x_i)}{p_{\theta_0(x_i)}}$ satisfies

$$\log L_{n,h} = Z_n^T h - \frac{1}{2} h^T I(\theta_0) h + r_n \text{ where } r_n \xrightarrow{\text{P}} 0,\ Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla l(\theta_0 | x_i).$$

– For some T_n of interest, write also

$$T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(x_i) + r_n \text{ where } r_n \xrightarrow{\text{P}} 0 \text{ for some function } f.$$

The joint convergence of such $(T_n, \log L_{n,h})$ typically comes from a bivariate Central Limit Theorem involving Z_n.

Example 6.3.2 (T-test)**.** Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$ where $p_\theta(x_i) = f(x_i - \theta),\ \theta \in \mathbb{R}$. Assume under $\theta = 0$,

$$\mathbb{E}_{\theta=0}[X_i] = \int x_i f(x_i) \mathrm{d}x_i = 0$$

$$Var_{\theta=0}[X_i] = \mathbb{E}_{\theta=0}\left[X_i^2\right] = \int x_i^2 f(x_i) \mathrm{d}x_i = \sigma^2 > 0.$$

$$x \cdot f(x) \to 0 \text{ as } x \to \pm\infty.$$

Consider the one-sided t-test of

$$H_0 : \theta = 0 \text{ v.s. } H_1 : \theta > 0.$$

which rejects for $T_n(X) > t_{n-1}^{(1-\alpha)}$ where

$$T_n(X) = \frac{\sqrt{n}\bar{X}}{S_n(X)}, \quad S_n(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Recall (4.1.1) and (4.3.2) $\sqrt{n}\bar{X} \xrightarrow{\text{d}} N(0, \sigma^2),\ S_n(X)^2 \xrightarrow{\text{P}} \sigma^2$ under $H_0 : \theta = 0$. So under $\theta = 0$,

$$T_n(X) = \frac{\sqrt{n}}{\sigma}\bar{X} + \sqrt{n}\bar{X}\left(\frac{1}{S_n} - \frac{1}{\sigma}\right) \text{ where } \sqrt{n}\bar{X}\left(\frac{1}{S_n} - \frac{1}{\sigma}\right) \xrightarrow{\text{P}} 0$$

Then an asymptotic level-α test rejects $H_0 : \theta = 0$ if $T_n(X) > z^{(1-\alpha)}$. And also $(T_n, \log L_{n,h})$ has the same limit under $H_0 : \theta = 0$ as

$$\left(\frac{\sqrt{n}}{\sigma}\bar{X},\ h \cdot Z_n - \frac{1}{2}\sigma_h^2\right) = \left(0,\ -\frac{1}{2}\sigma_h^2\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{X_i}{\sigma},\ l'(\theta_0 | X_i) \cdot h\right).$$

$$\text{where } \sigma_h^2 = h^T I(\theta_0) h,\ l'(\theta_0 | X_i) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_i - \theta)\big|_{\theta=0} = -\frac{f'(X_i)}{f(X_i)}.$$

So under $H_0 : \theta = 0$, by bivariate Central Limit Theorem

$$(T_n, \log L_{n,h}) \xrightarrow{\text{d}} N\left(\left(0, -\frac{1}{2}\sigma_h^2\right), \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_h^2 \end{pmatrix}\right).$$

$$\begin{aligned}
\text{where } \sigma_{1,2} &= \mathbb{E}_{\theta=0}\left[\frac{X_i}{\sigma} \cdot l'(\theta_0|X_i) \cdot h\right] \\
&= -\frac{h}{\sigma} \cdot \mathbb{E}_{\theta=0}\left[X_i \cdot \frac{f'(X_i)}{f(X_i)}\right] \\
&= -\frac{h}{\sigma} \cdot \int_{-\infty}^{\infty} x_i \cdot \frac{f'(x_i)}{f(x_i)} \cdot f(x_i)\mathrm{d}x \\
&= \frac{h}{\sigma} \cdot \int_{-\infty}^{\infty} f(x_i)\mathrm{d}x_i \text{ integrate by parts} \\
&= \frac{h}{\sigma}.
\end{aligned}$$

So by Le Cam's Third Lemma, under the alternative $\theta = \frac{h}{\sqrt{n}}$,

$$T_n \xrightarrow{\text{d}} N\left(\frac{h}{\sigma}, 1\right).$$

The asymptotic power under $\theta = \frac{h}{\sqrt{n}}$ as $n \to \infty$ is

$$\begin{aligned}
\mathbb{P}_{\theta=\frac{h}{\sqrt{n}}}[T_n > z^{(1-\alpha)}] &\to \mathbb{P}\left[N\left(\frac{h}{\sigma}, 1\right) > z^{1-\alpha}\right] \\
&= \mathbb{P}\left[Z + \frac{h}{\sigma} > z^{1-\alpha}\right] \text{ where } Z \sim N(0,1) \\
&= \tilde{\Phi}\left(z^{(1-\alpha)} - \frac{h}{\sigma}\right) \text{where } \tilde{\Phi}(x) = \mathbb{P}[Z > x] \text{ for } Z \sim N(0,1).
\end{aligned}$$

Compare: In Example 6.3.1, the asymptotic power for the likelihood ratio test (the most powerful test) is $\Phi\left(z^{(1-\alpha)} - \sigma_h\right)$ where

$$\sigma_h = \sqrt{h^T I(\theta_0) h} = h \cdot \sqrt{I(0)}.$$

By Cramer-Rao Lower Bound,

$$\frac{\sigma^2}{n} = Var(\bar{X}) \geq \frac{1}{n \cdot I(0)} \implies \sqrt{I(0)} \geq \frac{1}{\sigma}.$$

So the t-test is asmptotically most powerful if and only if $I(0) = \frac{1}{\sigma}$ (i.e., $\bar{X}$ is an asymptotically efficient estimator for θ in this model).

Example 6.3.3 (Sign test)**.** Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$ where $p_\theta(x_i) = f(x_i - \theta)$. Assume under $\theta = 0$,

$\mathbb{P}_{\theta=0}[X_i > 0] = \mathbb{P}_{\theta=0}[X_i < 0] = \frac{1}{2}$ (i.e., 0 is the median for x_i).

$f(x) \to 0$ as $x \to \pm\infty$.

Consider testing

$$H_0 : \theta = 0 \text{ v.s. } H_1 : \theta > 0.$$

using the sign statistic

$$S_n = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \mathbb{1}\{X_i > 0\} - \frac{1}{2} \right).$$

Under H_0, $\mathbb{1}\{X_i > 0\} \sim Bernoulli(\frac{1}{2})$ and the variance is $\frac{1}{4}$. So by Central Limit Theorem,

$$S_n \xrightarrow{\text{d}} N(0, \frac{1}{4}).$$

An asymptotic level-α test rejects $H_0 : \theta = 0$ if $S_n > \frac{1}{2} z^{(1-\alpha)}$.
Then the pair $(S_n, \log L_{n,h})$ has the same limit under $H_0 : \theta = 0$ as

$$\left(S_n,\ h \cdot Z_n - \frac{1}{2}\sigma_h^2 \right) = \left(0,\ -\frac{1}{2}\sigma_h^2 \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{1}\{X_i > 0\} - \frac{1}{2},\ l'(\theta_0 | X_i) \cdot h \right).$$

where $\sigma_h^2 = h^T I(\theta_0) h,\ l'(\theta_0 | X_i) = \dfrac{\text{d}}{\text{d}\theta} \log f(X_i - \theta)\big|_{\theta=0} = -\dfrac{f'(X_i)}{f(X_i)}$.

So under $H_0 : \theta = 0$, by bivariate Central Limit Theorem

$$(S_n, \log L_{n,h}) \xrightarrow{\text{d}} N\left(\left(0, -\frac{1}{2}\sigma_h^2 \right), \begin{pmatrix} \frac{1}{4} & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_h^2 \end{pmatrix} \right).$$

$$\begin{aligned}
\text{where } \sigma_{1,2} &= Cov_{\theta=0}\left[\mathbb{1}\{X_i > 0\} - \frac{1}{2}, l'(\theta_0 | X_i) \cdot h \right] \\
&= \mathbb{E}_{\theta=0}\left[\mathbb{1}\{X_i > 0\} \cdot l'(\theta_0 | X_i) \cdot h \right] \\
&= -h \cdot \mathbb{E}_{\theta=0}\left[\mathbb{1}\{X_i > 0\} \cdot \frac{f'(X_i)}{f(X_i)} \right] \\
&= -h \cdot \int_0^\infty \frac{f'(x_i)}{f(x_i)} \cdot f(x_i) \text{d}x \\
&= -h \cdot \int_0^\infty f'(x_i) \text{d}x_i \text{ integrate by parts} \\
&= h \cdot f(0).
\end{aligned}$$

So by Le Cam's Third Lemma, under the alternative $\theta = \frac{h}{\sqrt{n}}$,

$$S_n \xrightarrow{\text{d}} N\left(h \cdot f(0), \frac{1}{4} \right).$$

The asymptotic power under $\theta = \frac{h}{\sqrt{n}}$ as $n \to \infty$ is

$$\begin{aligned}
\mathbb{P}_{\theta=\frac{h}{\sqrt{n}}}[S_n > z^{(1-\alpha)}] &\to \mathbb{P}\left[N\left(h \cdot f(0), \frac{1}{4} \right) > \frac{1}{2} z^{1-\alpha} \right] \\
&= \mathbb{P}\left[h \cdot f(0) + \frac{1}{2} Z > \frac{1}{2} z^{1-\alpha} \right] \text{ where } Z \sim N(0,1) \\
&= \tilde{\Phi}\left(z^{(1-\alpha)} - 2 \cdot h f(0) \right) \text{where } \tilde{\Phi}(x) = \mathbb{P}[Z > x] \text{ for } Z \sim N(0,1).
\end{aligned}$$

Compare: In Example 6.3.1, the asymptotic power for the likelihood ratio test (the most powerful test) is $\Phi\left(z^{(1-\alpha)} - \sigma_h\right)$ where

$$\sigma_h = \sqrt{h^T I(\theta_0) h} = h \cdot \sqrt{I(0)}.$$

Fact: Sample median $\hat{\theta}_n$ of $x_1, x_2, \cdots, x_n$ is asymptotically normal around the true median θ.

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\text{d}} N\left(0, \frac{1}{4f^2(0)}\right).$$

By Cramer-Rao Lower Bound,

$$\frac{1}{n} \cdot \frac{1}{4f^2(0)} = Var(\hat{\theta}_n) \geq \frac{1}{n \cdot I(0)} \implies \sqrt{I(0)} \geq 2f(0).$$

So the sign test is asmptotically most powerful if and only if $\sqrt{I(0)} = 2f(0)$ (i.e., the sample median is an asymptotically efficient estimator for θ in this model).

Remark 6.3.2.

- The Neyman-Pearson test achieves asymptotic power $\tilde{\Phi}\left(z^{(1-\alpha)} - h^T I(\theta_0) h\right)$ but requires us to fix a particular local alternative h.
- The T-test and sign test do not use the knowledge of h, but only achieve this this best asymptotic power $\tilde{\Phi}\left(z^{(1-\alpha)} - \sqrt{h^T I(\theta_0) h}\right)$ in resticted settings.

6.4 Local Optimality in Testing

Recall: In finite-sample theory, by Theorem 3.2.2, there is a one-sided UMP test against $\theta > \theta_0$ for one-parameter exponential family models, or more generally only if the likelihood ratio $\prod_{i=1}^{n} \frac{p_{\theta_1}(x_i)}{p_{\theta_0}(x_i)}$ is monotonically increasing or decreasing in same fixed statistic $T(X)$ for all $\theta_1 > \theta_0$.

Question 6.4.1. To test $\theta = \theta_0 + \frac{h}{\sqrt{n}}$, is there a single test that achieves the optimal asymptotic power $\tilde{\Phi}\left(z^{(1-\alpha)} - h\sqrt{I(\theta_0)}\right)$ for all local alternatives $h > 0$?

Note 6.4.1. Such a test is called "locally asymptotically UMP".

Firstly, we consider testing the one-sided alternative.

Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \Omega \subseteq \mathbb{R}$. Test

$$H_0 : \theta = \theta_0 \text{ v.s. } \theta > \theta_0.$$

Answer 6.4.1. Rao's score test and Wald test.

- Rao's score test: By Theorem 6.2.1, under $H_0 : \theta = \theta_0$,

$$\log L_{n,h} \equiv \log \frac{\prod_{i=1}^{n} p_{\theta_0 + \frac{h}{\sqrt{n}}}(x_i)}{\prod_{i=1}^{n} p_{\theta_0}(x_i)} = h \cdot Z_n - \frac{1}{2} h^2 I(\theta_0) + r_n, \text{ where } r_n \xrightarrow{\text{P}} 0.$$

$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(\theta_0 | x_i)$ is the *score statistic* that doesn't depend on h.

Use Z_n as test statistic, under $H_0 : \theta = \theta_0$,

$$Z_n \xrightarrow{\text{d}} N(0, I(\theta_0)).$$

So an asymptotic level-α *score test* rejects $H_0 : \theta = \theta_0$ if $Z_n > \sqrt{I(\theta_0)} \cdot z^{(1-\alpha)}$.

Proposition 6.4.1. If the model is q.m.d. at $\theta = \theta_0$, then the score test achieves optimal power under $H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}$ for every fixed $h > 0$ as $n \to \infty$,

$$\mathbb{P}_{\theta_0 + \frac{h}{\sqrt{n}}} \left[Z_n > \sqrt{I(\theta_0)} \cdot z^{(1-\alpha)} \right] \to \tilde{\Phi}(z^{(1-\alpha)} - h\sqrt{I(\theta_0)})$$
$$\text{where } \tilde{\Phi}(x) = \mathbb{P}[Z > x] \text{ for } Z \sim N(0,1).$$

Proof. The joint limiting distribution of $(Z_n, \log L_{n,h})$ under θ_0 is the same as that of $\left(Z_n, h \cdot Z_n - \frac{1}{2}h^2 I(\theta_0)\right)$. As $n \to \infty$, Z_n and $\log L_{n,h}$ are perfectly correlated and has the degenerate "bivariate normal" distribution

$$\left(Z_n, h \cdot Z_n - \frac{1}{2}h^2 I(\theta_0) \right) \xrightarrow{\text{d}} N\left(\left(0, -\frac{1}{2}h^2 I(\theta_0) \right), \begin{pmatrix} I(\theta_0) & hI(\theta_0) \\ hI(\theta_0) & h^2 I(\theta_0) \end{pmatrix} \right).$$

So by Le Cam's Third Lemma, under $H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}$

$$Z_n \xrightarrow{\text{d}} N(hI(\theta_0), I(\theta_0)).$$

The asymptotic power under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$ as $n \to \infty$ is

$$\begin{aligned} \mathbb{P}_{\theta_0 + \frac{h}{\sqrt{n}}} \left[Z_n > \sqrt{I(\theta_0)} \cdot z^{(1-\alpha)} \right] &\to \mathbb{P}\left[N(hI(\theta_0), I(\theta_0)) > \sqrt{I(\theta_0)} \cdot z^{(1-\alpha)} \right] \\ &= \mathbb{P}\left[h \cdot I(\theta_0) + \sqrt{I(\theta_0)} \cdot Z > \sqrt{I(\theta_0)} \cdot z^{(1-\alpha)} \right] \\ &= \tilde{\Phi}\left(z^{(1-\alpha)} - h\sqrt{I(\theta_0)} \right). \end{aligned}$$

□

Note 6.4.2. If we want to test against *local* alternatives $\theta = \theta_0 + \frac{h}{\sqrt{n}}$ when n is large, the likelihood ratio is approximately monotonic in the score Z_n regardless of the details of the model. Hence, there exists an asymptotic UMP test against local alternatives in much greater generality than the existence of a UMP test for small n.

– Wald test: Based on the MLE $\hat{\theta}_n$, or more generally an efficient root of the likelihood equation

$$0 = \sum_{i=1}^{n} l'(\hat{\theta}_n | x_i).$$

Recall Theorem 5.4.2, under regularity assumptions, under $\theta = \theta_0$, as $n \to \infty$

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1} \cdot Z_n + r_n, \text{ where } r_n \xrightarrow{\text{P}} 0. \\ &\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N(0, I(\theta_0)^{-1}). \end{aligned} \tag{6.4.1}$$

The *Wald test* rejects $H_0 : \theta = \theta_0$ when $\sqrt{n}(\hat{\theta}_n - \theta_0) > \frac{1}{\sqrt{I(\theta_0)}} z^{(1-\alpha)}$.

Proposition 6.4.2. If the model is q.m.d. at θ_0 and in addition, (6.4.1) holds, then the Wald test achieves asymptotically optimal power under $H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}$ for every fixed $h > 0$ as $n \to \infty$,

$$\mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[\sqrt{n}(\hat{\theta}_n - \theta_0) > \frac{1}{\sqrt{I(\theta_0)}} z^{(1-\alpha)}\right] \to \tilde{\Phi}(z^{(1-\alpha)} - h\sqrt{I(\theta_0)})$$

where $\tilde{\Phi}(x) = \mathbb{P}[Z > x]$ for $Z \sim N(0,1)$.

Proof. Under (6.4.1),

$$\begin{aligned}
&\mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[\sqrt{n}(\hat{\theta}_n - \theta_0) > \frac{1}{\sqrt{I(\theta_0)}} z^{(1-\alpha)}\right] \\
&= \mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[I(\theta_0)^{-1} \cdot Z_n + r_n > \frac{1}{\sqrt{I(\theta_0)}} z^{(1-\alpha)}\right] \\
&= \mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[Z_n > \sqrt{I(\theta_0)} z^{(1-\alpha)} - I(\theta_0) \cdot r_n\right] \\
&\to \tilde{\Phi}\left(z^{(1-\alpha)} - h\sqrt{I(\theta_0)}\right) \text{ by Proposition 6.4.1.}
\end{aligned}$$

□

Secondly, we consider testing the two-sided alternative.

Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \Omega \subseteq \mathbb{R}$. Test

$$H_0 : \theta = \theta_0 \text{ v.s. } \theta \neq \theta_0.$$

Answer 6.4.2. There is no UMP test even in an asymptotic sense. However, reasonable tests are given by "symmetric" versions of the Rao's score test, Wald test. We can also use generalized likelihood ratio test alternatively.

– Rao's score test: reject $H_0 : \theta = \theta_0$ if

$$|Z_n| > \sqrt{I(\theta_0)} \cdot z^{\left(1-\frac{\alpha}{2}\right)} \iff Z_n^2 > I(\theta_0) \cdot \left(\mathcal{X}_1^2\right)^{(1-\alpha)}.$$

$\left(z^{(1-\frac{\alpha}{2})}\right)^2 = \left(\mathcal{X}_1^2\right)^{(1-\alpha)}$ are both $(1-\alpha)$-quantile of Z^2 when $Z \sim N(0,1)$.

– Wald test: reject $H_0 : \theta = \theta_0$ if

$$\sqrt{n}\left|\hat{\theta_n} - \theta_0\right| > \frac{1}{\sqrt{I(\theta_0)}} \cdot z^{\left(1-\frac{\alpha}{2}\right)} \iff n\left(\hat{\theta_n} - \theta_0\right)^2 > \frac{1}{I(\theta_0)} \cdot \left(\mathcal{X}_1^2\right)^{(1-\alpha)}.$$

Proposition 6.4.3. Under the preceding conditions, testing against the alternative $\theta = \theta_0 + \frac{h}{\sqrt{n}}$ (for either $h > 0$ or $h < 0$), these tests achieve asymptotic power $\tilde{F}\left(\left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right)$, where $\tilde{F}(x) = \mathbb{P}[Y \geq x]$ when Y has the non-central $\mathcal{X}^2$-distribution $\mathcal{X}_1'^2\left(h^2 I(\theta_0)\right)$ ($\mathcal{X}_1'^2(\lambda)$ is the law of Z^2 when $Z \sim N(\sqrt{\lambda}, 1)$).

Proof. Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$,

$$Z_n \xrightarrow{\mathrm{d}} N(hI(\theta_0), i(\theta_0)) = \sqrt{I(\theta_0)} \cdot N(h\sqrt{I(\theta_0)}, 1)$$
$$Z_n^2 \xrightarrow{\mathrm{d}} I(\theta_0) \cdot \mathcal{X}_1'^2\left(h^2 I(\theta_0)\right)$$
$$\mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[Z_n^2 > I(\theta_0) \cdot \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] = \mathbb{P}\left[Y \geq \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] = \tilde{F}\left(\left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right)$$
When $Y \sim \mathcal{X}_1'^2\left(h^2 I(\theta_0)\right)$.

Analysis of the Wald test is the same. □

– Generalized likelihood ratio test (GLRT): Let $l(\theta) = \sum_{i=1}^{n} l(\theta|x_i)$ be the log-likelihood, $\hat{\theta}_n$ be the MlE for θ. Reject $H_0 : \theta = \theta_0$ if

$$T_n \equiv 2 \cdot \left(l(\hat{\theta}_n) - l(\theta_0)\right) > (\mathcal{X}_1^2)^{(1-\alpha)}.$$

Proposition 6.4.4. If the model is q.m.d. at $\theta = \theta_0$ and in addition, (6.4.1) holds, then

$$\text{Under } H_0 : \theta = \theta_0, \quad \mathbb{P}_{\theta_0}\left[T_n > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] \to \alpha.$$
$$\text{Under } H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}, \quad \mathbb{P}_{\theta_0}\left[T_n > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] \to \tilde{F}\left(\left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right).$$

Proof sketch. Let $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$, so $\hat{\theta}_n = \theta_0 + \frac{\hat{h}_n}{\sqrt{n}}$. Then

$$T_n \equiv 2 \cdot \left(l(\hat{\theta}_n) - l(\theta_0)\right) = 2 \cdot \frac{\prod_{i=1}^{n} p_{\hat{\theta}_n}(x_i)}{\prod_{i=1}^{n} p_{\theta_0}(x_i)} = 2 \log L_{n,\hat{h}_n}.$$

Fix $c > 0$ and write

$$\mathbb{P}_{\theta_0}\left[2 \log L_{n,\hat{h}_n} > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] \leq \mathbb{P}_{\theta_0}\left[2 \log L_{n,\hat{h}_n} > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}, \left|\hat{h}_n\right| \leq c\right] + \mathbb{P}_{\theta_0}\left[\left|\hat{h}_n\right| > c\right].$$

Set $\epsilon_{n,c} = \sup_{h \in [-c,c]} \left|\log L_{n,h} - h \cdot Z_n + \frac{1}{2}h^2 I(\theta_0)\right|$.
For each fixed $h \in [-c, c]$, under $\theta = \theta_0$, we showed

$$\left|\log L_{n,h} - h \cdot Z_n + \frac{1}{2}h^2 I(\theta_0)\right| \xrightarrow{\mathrm{P}} 0.$$

The same proof shows, in fact, that if $h_n \to h$ as $n \to \infty$, then

$$\left|\log L_{n,h_n} - h_n \cdot Z_n + \frac{1}{2}h_n^2 I(\theta_0)\right| \xrightarrow{\mathrm{P}} 0 \quad (*).$$

This implies $\epsilon_{n,c} \xrightarrow{\mathrm{P}} 0$.

Proof. Suppose $\epsilon_{n,c} \xrightarrow{\mathrm{P}} 0$ not holds, then there exists $\delta > 0$ and a sequence $h_{n_1}, h_{n_2}, \cdots \in [-c, c]$ where for all $i = 1, 2, \cdots$

$$\mathbb{P}\left[\left|\log L_{n_i,h_{n_i}} - h_{n_i} \cdot Z_{n_i} + \frac{1}{2} h_{n_i}^2 I(\theta_0)\right| > \delta\right] > \delta \quad (**).$$

Since $[-c, c]$ is compact, there is a subsequence of $\{h_{n_i}\}_{i=1}^{\infty}$ that converges to some $h \in [-c, c]$. But $(**)$ holds along this subsequence contradicting $(*)$. $\square$

Under $\theta = \theta_0$,

- For one direction, note that if

$$2 \cdot \log L_{n,\hat{h}_n} > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}, \left|\hat{h}_n\right| \leq c.$$

Then

$$2 \cdot \hat{h}_n Z_n - \hat{h}_n^2 I(\theta_0) > \left(\mathcal{X}_1^2\right)^{(1-\alpha)} - 2\epsilon_{n,c}.$$

Under $(*)$

$$\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1} Z_n + r_n. \text{ where } r_n \xrightarrow{\mathrm{P}} 0.$$

So

$$2 \cdot \hat{h}_n Z_n - \hat{h}_n^2 I(\theta_0) = I(\theta_0)^{-1} \cdot Z_n^2 + \tilde{r}_n \text{ where } \tilde{r}_n \xrightarrow{\mathrm{P}} 0.$$

Then as $I(\theta_0)^{-1} \cdot Z_n^2 \xrightarrow{\mathrm{d}} \mathcal{X}_1^2$.

$$\begin{aligned}
&\mathbb{P}_{\theta_0}\left[2 \log L_{n,\hat{h}_n} > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}, \left|\hat{h}_n\right| \leq c\right] \\
&\leq \mathbb{P}_{\theta_0}\left[I(\theta_0)^{-1} \cdot Z_n^2 > \left(\mathcal{X}_1^2\right)^{(1-\alpha)} - \tilde{r}_n - 2\epsilon_{n,c}\right] \to \alpha.
\end{aligned}$$

Furthermore,

$$\mathbb{P}_{\theta_0}\left[\left|\hat{h}_n\right| > c\right] \to \mathbb{P}\left[|W| > c\right] \text{ where } W \sim N(0, I(\theta_0)^{-1}).$$

Taking $c \nearrow \infty$, we obtain $\mathbb{P}\left[|W| > c\right] \to 0$, hence

$$\lim_{n\to\infty} \sup \mathbb{P}_{\theta_0}\left[T_n > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] \leq \alpha.$$

- For the other directions, we have

$$\begin{aligned}
\mathbb{P}_{\theta_0}\left[T_n > \left(\mathcal{X}_1^2\right)^{(1\alpha)}\right] &\geq \mathbb{P}_{\theta_0}\left[2 \log L_{n,\hat{h}_n} > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}, \left|\hat{h}_n\right| \leq c\right] \\
&\geq \mathbb{P}_{\theta_0}\left[I(\theta_0)^{-1} \cdot Z_n^2 > \left(\mathcal{X}_1^2\right)^{(1-\alpha)} - \tilde{r}_n + 2\epsilon_{n,c}\right] \\
&\quad - \mathbb{P}_{\theta_0}\left[\left|\hat{h}_n\right| > c\right].
\end{aligned}$$

Similarly, we have

$$\lim_{n\to\infty} \sup \mathbb{P}_{\theta_0}\left[T_n > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] \geq \alpha.$$

Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$, the proof is the same, where instead we use

$$\mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[I(\theta_0)^{-1}Z_n > \left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right] \to \tilde{F}\left(\left(\mathcal{X}_1^2\right)^{(1-\alpha)}\right).$$
$$\mathbb{P}_{\theta_0+\frac{h}{\sqrt{n}}}\left[\left|\hat{h}_n\right| > c\right] \to \mathbb{P}[|W| > c] \text{ where } W \sim N(h, I(\theta_0)^{-1}).$$

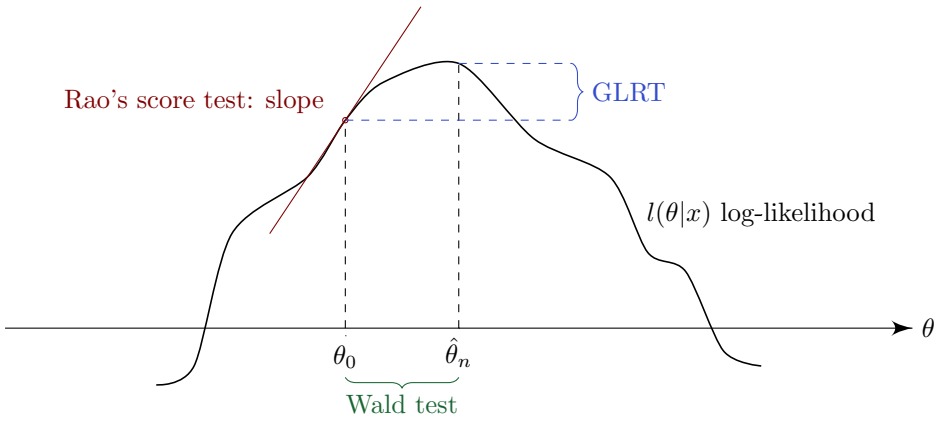
For fixed h, we still have $\mathbb{P}[|W| > c] \to 0$ as $c \nearrow \infty$. □

Remark 6.4.1. Under a quadratic approximation,

$$l(\theta|x) \approx l(\hat{\theta}_n|x) - \frac{1}{2}I(\theta_0)^{-1}\left(\theta - \hat{\theta}_n\right).$$

All three statistics contain the same information about how far $\hat{\theta}_n$ is from θ_0.

Comparison of Rao's score test, Wald test, GLRT



Thirdly, we test the two-sided alternative in higher dimensions.

Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \Omega \subseteq \mathbb{R}^k$. Test

$$H_0 : \theta = \theta_0 \text{ v.s. } \theta \neq \theta_0.$$

Answer 6.4.3. There are "chi-squared analogues" of these tests:

– Rao's score test: $Z_n^T I(\theta_0)^{-1} Z_n \in \mathbb{R}^k$.

- $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla l(\theta_0|x_i)$.
- Under $\theta = \theta_0$, $Z_n \xrightarrow{\text{d}} N(0, I(\theta_0))$. Then

$$Z_n^T I(\theta_0)^{-1} Z_n \xrightarrow{\text{d}} \mathcal{X}_k^2.$$

- Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$, $Z_n \xrightarrow{\text{d}} N(I(\theta_n)^T h, I(\theta_0))$. Then

$$Z_n^T I(\theta_0)^{-1} Z_n \xrightarrow{\text{d}} \mathcal{X}_k'^2\left(h^T I(\theta_0) h\right).$$

– Wald test: $n(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0) \in \mathbb{R}^k$.

  ◦ Under $\theta = \theta_0$, $\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1} \cdot Z_n + r_n$ where $r_n \xrightarrow{\text{P}} 0$. So $n(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0) = Z_n^T I(\theta_0)^{-1} Z_n + r_n'$, where $r_n' \xrightarrow{\text{P}} 0$. Then

$$n(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} \mathcal{X}_k^2.$$

  ◦ Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$, we can show that

$$n(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} \mathcal{X}_k'^2 \left(h^T I(\theta_0) h\right).$$

– GLRT: $2(l(\hat{\theta}_n - l(\theta_0))) \in \mathbb{R}$.

  ◦ $T_n = 2(l(\hat{\theta}_n - l(\theta_0))) = 2\sum_{i=1}^n l(\hat{\theta}_n | x_i) - 2\sum_{i=1}^n l(\hat{\theta}_n | x_i)$.

  ◦ Under $\theta = \theta_0$, similarly, we can show this equivalent to $Z_n^T I(\theta_0) Z_n$. Then

$$2(l(\hat{\theta}_n - l(\theta_0))) \xrightarrow{\text{d}} \mathcal{X}_k^2.$$

  ◦ Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$, we can show that

$$2(l(\hat{\theta}_n - l(\theta_0))) \xrightarrow{\text{d}} \mathcal{X}_k'^2 \left(h^T I(\theta_0) h\right).$$

Lastly, we test the two-sided alternative in with nuissance parameters.

Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta = (\alpha, \beta) \in \Omega \subseteq \mathbb{R}^k$, $\alpha \in \mathbb{R}^j$, $\beta \in \mathbb{R}^{k-j}$. Test

$$H_0 : \alpha = \alpha_0 \text{ v.s. } \alpha \neq \alpha_0.$$

– Rao's score test: Write $I(\theta) = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\beta} \\ I_{\beta\alpha} & I_{\alpha\beta} \end{pmatrix}$. Fix $a = \alpha_0$, compute MLE $\hat{\beta}_0$ for β in this submodel. Use the score statistic $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\alpha l(\alpha_0, \hat{\beta}_0)$. By Taylor expansion,

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\nabla_\alpha l(\alpha_0, \beta_0) + \nabla_{\alpha,\beta}^2 l(\alpha_0, \beta_0)(\hat{\beta}_0 - \beta_0)\right).$$

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) = I_{\beta\beta}^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\beta l(\alpha_0, \beta_0).$$

By Central Limit Theorem, under $H_0 : \alpha = \alpha_0$,

$$Z_n \xrightarrow{\text{d}} N\left(0, \check{I}_{\alpha\alpha}\right). \quad \check{I}_{\alpha\alpha} \text{ is different from } I_{\alpha\alpha}.$$

Then,

$$Z_n^T \ \check{I}_{\alpha\alpha}^{-1} \ Z_n \xrightarrow{\text{d}} \mathcal{X}_j^2.$$

– Wald test: Compute the MLE $(\hat{\alpha}, \hat{\beta})$. And test based on $\sqrt{n}\,(\hat{\alpha} - \alpha_0)$. Under $H_0 : \alpha = \alpha_0$, $\sqrt{n}\left(\hat{\alpha} - \alpha_0, \hat{\beta} - \beta_0\right) \xrightarrow{\text{d}} N\left(0, I(\alpha_0, \beta_0)^{-1}\right)$. So

$$\sqrt{n}\,(\hat{\alpha} - \alpha_0) \xrightarrow{\text{d}} N\left(0, \left[I(\alpha_0, \beta_0)^{-1}\right]_{\alpha\alpha}\right) \equiv N\left(0, \tilde{I}_{\alpha\alpha}^{-1}\right).$$
$$\tilde{I}_{\alpha\alpha}^{-1} = \left(I_{\alpha\alpha} - I_{\alpha\beta} I_{\beta\beta}^{-1} I(\beta\alpha)\right)^{-1}.$$

Then
$$n\,(\hat{\alpha} - \alpha_0)^T\, \tilde{I}_{\alpha\alpha}\,(\hat{\alpha} - \alpha_0) \xrightarrow{\text{d}} \mathcal{X}_j^2.$$

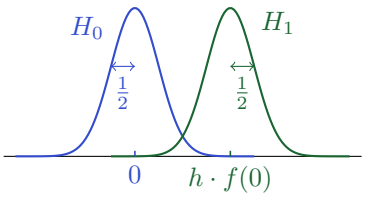
– GLRT: $T_n = 2\sum_{i=1}^{n} l(\hat{\alpha}, \hat{\beta}|x_i) - 2\sum_{i=1}^{n} l(\alpha_0, \hat{\beta}_0|x_i)$. Check by Taylor expansion that, under $H_0 : \alpha = \alpha_0$,
$$T_n \xrightarrow{\text{d}} \mathcal{X}_j^2.$$

6.5 The Limiting Normal Experiment

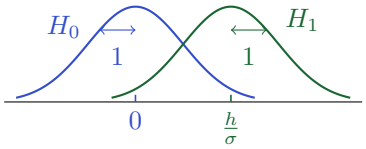
Recall: Suppose $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $p_\theta(x_i) = f(x_i - \theta)$, $\theta \in \mathbb{R}$. Suppose $\theta_0 = 0$ is the mean and median of f. We had the following asymptotic behavior of test statistics under

$$H_0 : \theta = 0 \text{ v.s. } H_1 : \theta = \frac{h}{\sqrt{n}}.$$

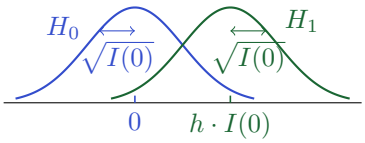
Sign statistic



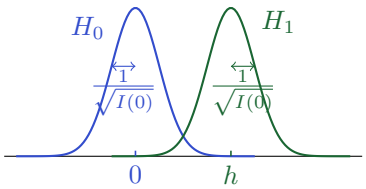
t-statistic



Rao's score statistic



Wald statistic



Consider the following separate statistical "experiment":

Observe a *single* normal $X \sim N(h, I(\theta_0)^{-1}) \in \mathbb{R}^k$, which is a Gaussian sequence model with known covariance $I(\theta_0)^{-1}$ and unknown mean h. This is called the *limiting normal experiment*. Then the log-likelihood ratio of this experiment is

$$\log \frac{p_h(x)}{p_0(x)} = \log \frac{\exp\left(-\frac{1}{2}(x-h)^T I(\theta_0)(x-h)\right)}{\exp\left(-\frac{1}{2}x^T I(\theta_0)x\right)} = h^T I(\theta_0)x - \frac{1}{2}h^T I(\theta_0)h.$$

Let $Z = I(\theta_0)X \sim N(I(\theta_0)h, I(\theta_0))$, we have

$$\log \frac{p_h(x)}{p_0(x)} = h^T Z - \frac{1}{2} h^T I(\theta_0) h \sim N\left(\frac{1}{2} h^T I(\theta_0) h, h^T I(\theta_0) h\right).$$

The behavior of score and likelihood ratio statistics of this simple Gaussian model exactly matches the asymptotic distributions of Z_n and $\log L_{n,h}$ under $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta + \frac{h}{\sqrt{n}}}$.

Informally: For $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta+\frac{h}{\sqrt{n}}}$ and large n, the amount of information we can learn about the local parameter h is equivalent to the amount of information we can learn about h upon observing a single normal $X \sim N(h, I(\theta_0)^{-1})$.

More formally:

Theorem 6.5.1. Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta+\frac{h}{\sqrt{n}}}$, where $\theta \in \Omega \subseteq \mathbb{R}^k$. Suppose the model is q.m.d. at θ_0. Let $T_n(X_1, X_2, \cdots, X_n)$ be any statistic (possibly depending on θ_0, but not on h), such that under $\theta_0 + \frac{h}{\sqrt{n}}$ for every fixed h, we have $T_n \xrightarrow{\text{d}} \mathcal{L}_h$ for some limit distribution $\mathcal{L}$. Then there exists a randomized statistic $T(X, U)$ in the limit experiment $X \sim N(h, I(\theta_0)^{-1})$, where $U \sim Uniform([0,1])$ is an independent source of randomness (also possibly depending on θ_0, but not on h), such that $T(X, U) \sim \mathcal{L}_h$.

Remark 6.5.1. In the previous examples:

- $T_n(X_1, X_2, \cdots, X_n) =$ Rao's score statistic $\iff T(X) = I(0) \cdot X$.
- $T_n(X_1, X_2, \cdots, X_n) =$ Wald statistic $\iff T(X) = X$.
- $T_n(X_1, X_2, \cdots, X_n) =$ Sign statistic $\iff T(X, Y) = f(0) \cdot X + \sqrt{\frac{1}{4} - \frac{f(0)^2}{I(0)}} \cdot Y$.
- $T_n(X_1, X_2, \cdots, X_n) =$ t-statistic $\iff T(X, Y) = \frac{1}{\sigma} \cdot X + \sqrt{1 - \frac{1}{\sigma^2 I(0)}} \cdot Y$.

Here, $Y \sim N(0,1)$ is an independent souce of randomness and we can generate Y as $Y = \Phi^{-1}(U)$. So $\mathbb{P}[Y \le y] = \mathbb{P}[U \le \Phi(y)] = \Phi(y)$. Extra randomness are not needed if $\frac{1}{4} - \frac{f(0)^2}{I(0)} = 0$ (sample median is efficient) or $1 - \frac{1}{\sigma^2 I(0)} = 0$ (sample mean is efficient).

To prove Theorem 6.5.1: Recall $\{Y_n\}$ are *bounded in probability* if, for any $\epsilon > 0$, there exists $M = M(\epsilon) > 0$ such that

$$\mathbb{P}[\|Y_n\| > M] < \epsilon \text{ for all } n.$$

If $Y_n \xrightarrow{\text{d}} Y$ for some limit Y, then $\{Y_n\}$ are bounded in probability. The following is a partial converse.

Lemma 6.5.1 (Prohorov's Theorem)**.** If $\{Y_n\}$ are bounded in probability, then there is a subsequence $Y_{n_1}, Y_{n_2}, \cdots$ that converges in distribution to a limit Y.

Example 6.5.1. Let $Y_n \sim N(\theta_n, 1)$. Suppose $\theta_n \in [-1, 1]$ for all n. Then $\{Y_n\}$ are bounded in probability: For any $\epsilon > 0$,

$$\mathbb{P}[|Y_n| > M] \le 2 \cdot \mathbb{P}[N(0,1) > M - 1] < \epsilon \text{ for large enough } M.$$

$\{Y_n\}$ may not converge in dsitribution. Take $\theta_n = 1$ for odd n, and $\theta_n = -1$ for even n. But by completeness of $[-1, 1]$, there is a subsequence $\theta_{n_1}, \theta_{n_2}, \cdots$ which converges to $\theta \in [-1, 1]$. Along this sequence. $Y_{n_1}, Y_{n_2}, \cdots \xrightarrow{\text{d}} N(\theta, 1)$.

Proof of Theorem 6.5.1. Let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla l(\theta_0 | x_i)$ be the score.

– Under $\theta = \theta_0$ $(h = 0)$: We know that T_n and Z_n marginally converge in distribution as $n \to \infty$, so (T_n, X_n) are bounded in probability. By Prohorov's Theorem, there is a subsequence $n_1, n_2, \cdots$ converges in bivaraite law:

$$(T_n, Z_n) \xrightarrow{\text{d}} (S, Z) \text{ where } S \sim \mathcal{L}_0 \text{ and } Z \sim N(0, I(\theta_0)).$$

Claim 6.5.1. Set $X = I(\theta_0)^{-1} Z$, so $X \sim N(0, I(\theta_0)^{-1})$. Then

$$(S, Z) \overset{\text{L}}{=} (T(X, U), I(\theta_0) X)$$

for some randomized statistic T.

Proof. Let $U \sim Uniform([0, 1])$ and F_x be the CDF of S conditional on $X = x$ for fixed x. Set $T(x, U) = F_x^{-1}(U)$. Then the conditional law of T given $X = x$ is the same as that of S.

$$\mathbb{P}[T(x, U) \le t] = \mathbb{P}[U \le F_x(t)] = F_x(t).$$

This holds for all x, so

$$(S, X) \overset{\text{L}}{=} (T(X, U), X) \implies (S, Z) \overset{\text{L}}{=} (T(X, U), I(\theta_0) X)\,.$$

Note: This assumes $T_n, S \in \mathbb{R}$. A similar argument applies for $T_n, S \in \mathbb{R}^j$. □

So we've represented $\mathcal{L}_0$ as $T(X, U)$ where $X \sim N(0, I(\theta_0)^{-1})$.

– Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$ $(h \neq 0)$:

Claim 6.5.2. Denote $\tilde{X} \sim N(h, I(\theta_0)^{-1})$. Under the above subsequence,

$$T_n \xrightarrow{\text{d}} T\left(\tilde{X}, U\right).$$

Proof. Under $\theta = \theta_0$ $(h = 0)$,

$$\log L_{n,h} = h^T Z_n - \frac{1}{2} h^T I(\theta_0) h + r_n \text{ where } r_n \xrightarrow{\text{P}} 0.$$

By Le Cam's Third Lemma, under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$,

$$(T_n, \log L_{n,h}) \xrightarrow{\text{d}} (\tilde{S}, \tilde{W}) \text{ where } \mathbb{P}\left[\left(\tilde{S}, \tilde{W}\right) \in B\right] = \mathbb{E}\left[\mathbb{1}\{(S, W) \in B\} \cdot e^W\right].$$

Here, $W = h^T Z - \frac{1}{2} h^T I(\theta_0) h$ is the limit of $\log L_{n,h}$, (S, Z) is the above limit of (T_n, Z_n) under $\theta = \theta_0$.

Set $X = I(\theta_0)^{-1}Z \sim N\left(0, I(\theta_0)^{-1}\right)$ and apply

$$(S, Z) \overset{\mathrm{L}}{=} (T(X, U), I(\theta_0)X)\,.$$

Then

$$\mathbb{P}\left[\left(\tilde{S}, \tilde{W}\right) \in B\right] =$$
$$\mathbb{E}\left[\mathbb{1}\{\left(T(X, U), h^T I(\theta_0)X - \frac{1}{2}h^T I(\theta_0)h\right) \in B\} \cdot e^{h^T I(\theta_0)X - \frac{1}{2}h^T I(\theta_0)h}\right].$$

Since $h^T I(\theta_0)X - \frac{1}{2}h^T I(\theta_0)h$ is the log-likelihood ratio between $\tilde{X} \sim N\left(h, I(\theta_0)^{-1}\right)$ and $X \sim \left(0, I(\theta_0)^{-1}\right)$, we have equivalently

$$\mathbb{P}\left[\left(\tilde{S}, \tilde{W}\right) \in B\right] =$$
$$\mathbb{E}_{\tilde{X} \sim N(h, I(\theta_0)^{-1})}\left[\mathbb{1}\{\left(T(\tilde{X}, U), h^T I(\theta_0)\tilde{X} - \frac{1}{2}h^T I(\theta_0)h\right) \in B\}\right].$$

So $\tilde{S} \overset{\mathrm{L}}{=} T(\tilde{X}, U)$, then under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$

$$T_n \xrightarrow{\mathrm{d}} T(\tilde{X}, U) \text{ where } \tilde{X} \sim N\left(h, I(\theta_0)^{-1}\right).$$

□

Since the theorem assumes $T_n \xrightarrow{\mathrm{d}} \mathcal{L}_n$ under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$ for each fixed h. Then $T(\tilde{X}, U) \sim \mathcal{L}_h$.

□

6.6 Asymptotic Optimality in Estimation

The limiting normal experiment can also be implied to estimation: Let $\delta_n\left(X_1, X_2, \cdots, X_n\right)$ be an estimate of $\theta \in \Omega \subseteq \mathbb{R}^k$, where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$. Fix θ_0, and suppose under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$,

$$\sqrt{n}\left(\delta_n(X_1, X_2, \cdots, X_n) - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \xrightarrow{\mathrm{d}} \tilde{\mathcal{L}}_h \quad \text{for each } h.$$

If the model is q.m.d. at θ_0, there is a randomized estimator $\delta(X, U)$ in the limit experiment $X \sim N\left(h, I(\theta_0)^{-1}\right)$ that has this dsitribution $\tilde{\mathcal{L}}_h$. Under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$, the rescaled estimation error is

$$\sqrt{n}\left(\delta_n(X_1, X_2, \cdots, X_n) - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) = \sqrt{n}\left(\delta_n(X_1, X_2, \cdots, X_n) - \theta_0\right) - h$$
$$\xrightarrow{\mathrm{d}} \delta(X, U) - h.$$

Here, $\delta(X, U) - h$ is the error of $\delta(X, U)$ as an estimator of h. We expect under squared-error loss,

$$\mathbb{E}_{\theta_0 + \frac{h}{\sqrt{n}}}\left[\left\|\sqrt{n}\left(\delta_n(X_1, X_2, \cdots, X_n) - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right)\right\|^2\right] \to \mathbb{E}\left[\|\delta(X, U) - h\|^2\right].$$

So

$$n \cdot R\left(\theta_0 + \frac{h}{\sqrt{n}}, \delta_n\right) \to \mathbb{E}\left[\|\delta(X,U) - h\|^2\right].$$

The asymptotic risk achievable by an estimate δ_n is $\frac{1}{n}$ times the risk achievable by an estimate of h in limiting experiment $X \sim N\left(h, I(\theta_0)^{-1}\right)$.

Example 6.6.1. Let $\delta_n(X_1, X_2, \cdots, X_n) = \hat{\theta}_n$ be the MLE. suppose

$$\hat{\theta}_n = I(\theta_0)^{-1} \cdot Z_n + r_n \quad \text{where } Z_n \text{ is the score and } r_n \xrightarrow{\text{P}} 0.$$

From the Wald test, under $\theta = \theta_0 + \frac{h}{\sqrt{n}}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N\left(h, I(\theta_0)^{-1}\right).$$

So

$$\sqrt{n}\left(\hat{\theta}_n - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) = \sqrt{n}(\hat{\theta}_n - \theta_0) - h \xrightarrow{\text{d}} N(0, I(\theta_0)^{-1}) \equiv \tilde{\mathcal{L}}_n.$$

This is matched by $\delta(X,U) = X$ in equivalent $X \sim N(h, I(\theta_0)^{-1})$. Under regularity conditions, we have

$$n \cdot R\left(\theta_0 + \frac{h}{\sqrt{n}}, \hat{\theta}_n\right) \to \mathbb{E}\left[\|X - h\|^2\right] = \mathbb{E}\left[\|W\|^2\right] \text{ where } W \sim N\left(0, I(\theta_0)^{-1}\right).$$

Question 6.6.1. Why not just assume $\sqrt{n}(\delta_n - \theta) \xrightarrow{\text{i.i.d.}} \mathcal{L}_\theta$ under every fixed $\theta \in \Omega$, and study the limits $\mathcal{L}_\theta$?

Example 6.6.2 (Hodges' "super efficient estimator")**.** Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$. Let $\delta_n(X_1, X_2, \cdots, X_n) = \begin{cases} \bar{X} & \left|\bar{X}\right| \geq n^{-\frac{1}{4}} \\ 0 & \left|\bar{X}\right| < n^{-\frac{1}{4}}. \end{cases}$

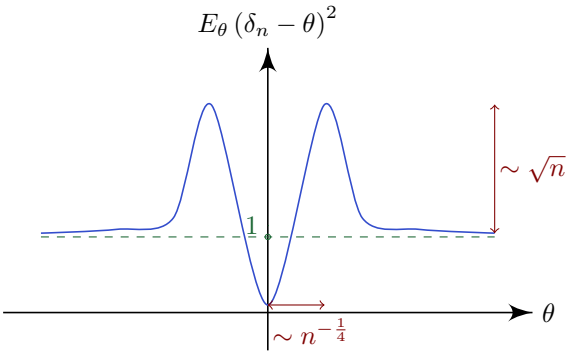
– At $\theta = 0$: $\bar{X} \sim N(0, \frac{1}{n})$, so $\mathbb{P}_{\theta=0}\left[\left|\bar{X}\right| < n^{-\frac{1}{4}}\right] \to 1$. Then

$$\mathbb{P}_{\theta=0}\left[\delta_n = 0\right] \to 1 \text{ and } \mathbb{E}_{\theta=0}\left[n^2(\delta_n - 0)^2\right] \to 0.$$

– For any fixed $\theta \neq 0$: $\bar{X} \sim N(\theta, \frac{1}{n})$, so $\mathbb{P}_\theta\left[\left|\bar{X}\right| < n^{-\frac{1}{4}}\right] \to 0$. Then

$$\mathbb{P}_\theta\left[\delta_n = \bar{X}\right] \to 1 \text{ and } \mathbb{E}_\theta\left[n^2(\delta_n - \theta)^2\right] \to 1.$$

We might conclude that $\delta(X_1, X_2, \cdots, X_n)$ "asymptotically dominates" the MLE $\bar{X}$. But this is misleading: the normalized risk $n \cdot (\delta_n - \theta)^2$ looks like



This is *much* larger than risk 1 of $\bar{X}$ at $\theta \approx n^{-\frac{1}{4}}$, but we don't see this under pointwise asymptotics. If we consider local alternative to $\theta = 0$, we see these bumps. Under a local alternative $\theta_n = \frac{h}{\sqrt{n}}$ at $\theta_0 = 0$,

$$\mathbb{P}_{\theta_n = \frac{h}{\sqrt{n}}} \left[\left| \bar{X} \right|_n < n^{-\frac{1}{4}} \right] \to 1.$$

So we estimate 0 with high probability

$$\mathbb{E}_{\theta_n = \frac{h}{\sqrt{n}}} \left[n \left(\delta_n - \frac{h}{\sqrt{n}} \right)^2 \right] \approx \mathbb{E}_{\theta_n = \frac{h}{\sqrt{n}}} \left[n \left(\frac{h}{\sqrt{n}} - 0 \right)^2 \right] = h^2.$$

The maximum limiting risk over a "local window" $h \in [-M, M]$ is M^2, which increases without bound as $M \to \infty$.

Theorem 6.6.1 (Hájek–Le Cam Local Asymptotic Minimax Theorem)**.** Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, $\theta \in \Omega \subseteq \mathbb{R}$. Suppose the model is q.m.d. at θ_0. Let δ_n be any estimator of θ. Then

$$\lim_{n \to \infty} \sup_{h \in [-M, M]} \mathbb{E}_{\theta_0 + \frac{h}{\sqrt{n}}} \left[n \cdot \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right)^2 \right] \geq \frac{1}{I(\theta_0)} - \epsilon(M) \tag{6.6.1}$$

where $\epsilon(M) \to 0$ as $M \nearrow \infty$.

Remark 6.6.1. Thus the Cramer-Rao Lower Bound represents a fundamental lower bound to the squared-error risk of any estimator, in a local minimax sense around θ_0, for each fixed $\theta_0 \in \mathbb{R}$.

Proof.

(i) Consider the limiting normal experiment $X \sim N(h, \frac{1}{I(\theta_0)})$. For estimating $h \in \mathbb{R}$, X is minimax, with constant risk

$$\mathbb{E}_h \left[(X - h)^2 \right] = \frac{1}{I(\theta_0)}.$$

Recall: take a prior $h \sim n(0, \tau^2)$. As $\tau \to \infty$, the Bayes risk increases to $\frac{1}{I(\theta_0)}$. This provides a lower bound to the minimax risk, so X is minimax.

Exercise: For estimating $h \in [-M, M]$,

$$\text{Minimax risk } \geq \frac{1}{I(\theta_0)} - \epsilon(M) \text{ for some } \epsilon(M) \to 0 \text{ as } M \nearrow \infty.$$

(ii) Consider any estimator δ_n of θ in the original model $\theta = \theta_0$. First suppose

$$\sqrt{n} \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right) \xrightarrow{\text{d}} \tilde{\mathcal{L}}_h \text{ under } \theta = \theta_0 + \frac{h}{\sqrt{n}} \text{ for each fixed } h.$$

Then there exists $\delta(X, U)$ in limiting experiment $X \sim N\left(h, \frac{1}{I(\theta_0)}\right)$ such that

$$\delta(X, U) - h \sim \tilde{\mathcal{L}}_h \implies \sqrt{n} \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right) \xrightarrow{\text{d}} \delta(X, U) - h.$$

Apply Portmanteau Lemma and $x \to x^2$ is non-negative, continuous,

$$\lim_{n\to\infty} \inf \mathbb{E}_{\theta_0+\frac{h}{\sqrt{n}}} \left[n \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right)^2 \right] \geq \mathbb{E}_h \left[(\delta(X,U) - h)^2 \right].$$

Hence,

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{h\in[-M,M]} \mathbb{E}_{\theta_0+\frac{h}{\sqrt{n}}} \left[n \cdot \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right)^2 \right] \quad (*) \\
&\geq \lim_{n\to\infty} \inf \mathbb{E}_{\theta_0+\frac{h}{\sqrt{n}}} \left[n \cdot \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right)^2 \right] \text{ for each fixed } h \in [-M, M] \\
&\geq \mathbb{E}_h \left[(\delta(X,U) - h)^2 \right]
\end{aligned}$$

This holdes for all $h \in [-M, M]$, so by (i)

$$(*) \geq \sup_{h\in[-M,M]} \mathbb{E}_h \left[(\delta(X,U) - h)^2 \right] \geq \frac{1}{I(\theta_0)} - \epsilon(M).$$

(iii) To remove the assumption in (ii),

$$\sqrt{n} \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right) \xrightarrow{\text{d}} \tilde{\mathcal{L}}_h \text{ under } \theta = \theta_0 + \frac{h}{\sqrt{n}} \text{ for each fixed } h.$$

Let $n_1, n_2, n_3, \cdots$ be subsequence where we attain

$$C \equiv \lim_{n\to\infty} \sup_{h\in[-M,M]} \mathbb{E}_{\theta_0+\frac{h}{\sqrt{n}}} \left[n \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right)^2 \right].$$

Under $\theta = \theta_0 (h = 0)$: For all large k, by Markov's Inequality,

$$\mathbb{E}_{\theta_0} \left[n_k \left(\delta_{n_k} - \theta_0 \right)^2 \right] \leq C + 1 \implies \mathbb{P}_{\theta_0} \left[\sqrt{n_k} \left| \delta_{n_k} - \theta_0 \right| > B \right] \leq \frac{C+1}{B^2}.$$

So $\sqrt{n_k}(\delta_{n_k} - \theta_0)$ is bounded in probability. By Prohorov's Theorem, there exists a sub-subsequence where

$\left(\sqrt{n}(\delta_n, \theta_0), Z_n \right) \xrightarrow{\text{d}} (\delta(X,U), I(\theta_0)X)$ for some matching estimator $\delta(X,U)$.

Mimicking proof of preceding theorem, along this sub-subsequence, under $\theta_0 + \frac{h}{\sqrt{n}}$ for each fixed h,

$$\sqrt{n} \left(\delta_n - \theta_0 \right) \xrightarrow{d} \delta(X,U) \text{ for } X \sim N\left(h, \frac{1}{I(\theta_0)} \right).$$

So

$$\sqrt{n} \left(\delta_n - \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right) \xrightarrow{\text{d}} \delta(X,U) - h.$$

Applying (ii) to this sub-subsequence, we have

$$C \geq \frac{1}{I(\theta_0)} - \epsilon(M).$$

□

Index