**STAT 581: Advanced Theory of Statistical Inference I** **Autumn 2026**

# Lecture 4: Hypothesis tests in likelihood models

*Instructor: Yen-Chi Chen*

⊙ We thank previous instructors: Jon Wellner, Alex Luedtke, Fang Han, and Andrea Rotnitzky.

⊙ Some of this lecture notes are based on the following book:

> [van der Vaart] Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

In particular, Chapter 6 and 7 are useful references.

## 4.1 Wald, score, and likelihood ratio tests

Suppose we have a random sample $X_1, \cdots, X_n \sim P_\theta$, where $P_\theta$ is a parametric model indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^d$. Let $p_\theta$ be the PDF/PMF of $P_\theta$.

We consider a common hypothesis testing problem that

$$H_0 : \theta \in \Theta_0 \subset \Theta \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1 \equiv \Theta \backslash \Theta_0.$$

Our decision on rejecting the null or not can be written as a function

$$\phi(X_1, \cdots, X_n) \in [0, 1]$$

such that $\phi(X_1, \cdots, X_n)$ is the probability of rejecting $H_0$. $\phi(X_1, \cdots, X_n) = 0$ means we do not reject and $\phi(X_1, \cdots, X_n) = 1$ is that we always reject. Such function $\phi(\cdot)$ is called a *test function*.

For a given parameter $\theta$, the *power function* is

$$\pi_n(\theta) = \mathbb{E}_\theta(\phi(X_1, \cdots, X_n)),$$

where $\mathbb{E}_\theta(\cdot)$ means that we assume $X_1, \cdots, X_n$ are IID from $P_\theta$. The power function is the probability of rejecting $H_0$ given a parameter. The expectation is over two things: the random sample $X_1, \cdots, X_n$ and the random decision of $\phi(X_1, \cdots, X_n)$ if $\phi(X_1, \cdots, X_n)$ is not 1 or 0.

### 4.1.1 Level-$\alpha$ test

The *type-1 error* is the probability of falsely reject the null. It is formally defined as $\sup_{\theta \in \Theta} \pi_n(\theta)$. In hypothesis test, controlling the type-1 error to be $\alpha$ means that

$$\sup_{\theta \in \Theta} \pi_n(\theta) \leq \alpha. \tag{4.1}$$

While we have asymptotic theory for many statistics, finite sample behavior is often difficult to control. Therefore, controlling type-1 error at *any* $n$ is very hard. So we often relax it to be asymptotic type-1 error

control. this leads to two possible options:

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \pi_n(\theta) \le \alpha, \tag{4.2}$$

$$\limsup_{n\to\infty} \pi_n(\theta) \le \alpha \qquad \text{for each } \theta \in \Theta. \tag{4.3}$$

Clearly, equation (4.2) is a stronger requirement than equation (4.3) since the former is a uniform result while the latter is a pointwise result. However, we have not yet learned the the required tool for showing equation (4.2)[1] so we will focus on equation (4.3).

If the test satisfies equation (4.1), it is called an *exact* level-$\alpha$ test. If the test satisfies equation (4.3), it is called an *asymptotic* level-$\alpha$ test.

### 4.1.2 Null hypothesis: simple cases

There are two common ways that people have used to describe null hypothesis when it only involves linear constraints. The first way is to express the null hypothesis as

$$H_0 : A\theta = c$$

for a given matrix $A \in \mathbb{R}^{k\times d}$ and $c \in \mathbb{R}^k$. This is the original form of the linear constraints.

The second way is to assume that we have do some simple transformations on $\theta$ so that the parameter is split into $\theta = (\psi, \eta)$ such that

$$H_0 : \psi = 0 \tag{4.4}$$

with $\psi \in \mathbb{R}^k$ and $\eta \in \mathbb{R}^{d-k}$. For the rest of this note, we will consider the null hypothesis in the form of equation (4.4) since it offers a very simple and elegant expression.

When $k = d$, the null hypothesis space $\Theta_n$ reduces to a singleton, and this scenario is called a simple null hypothesis. When $k < d$, the null hypothesis space $\Theta_n$ many elements so it is called a composite null hypothesis. Note that $k < d$ implies that $\Theta_n$ is in fact a $d - k$ dimensional linear subspace.

**Example 4.1 (Moment constraint of a 2-Gaussian mixture model)** *Suppose we assume that our data $X_1, \cdots, X_n \in \mathbb{R}$ is from a 2-Gaussian mixture model (2-GMM) such that the PDF is*

$$p_\theta(x) = 0.5 \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + 0.5 \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}.$$

*In this case, the parameter space is $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \in \mathbb{R}^2 \times \mathbb{R}^2_{>0} \subset \mathbb{R}^4$. Note that the proportion of the two components are assumed to be the same otherwise it could be another parameter.*

*Suppose our null hypothesis is that $\mathbb{E}(X_1) = 0$, i.e.,*

$$\mu_1 + \mu_2 = 0,$$

*so it is a linear constraint. In this case, $A = (1, 1, 0, 0)^T$ and $c = 0$ if we take linear constraint form. Or we can reparametrize the model as*

$$\theta = \left( \underbrace{\mu_1 + \mu_2}_{\psi}, \underbrace{\mu_1 - \mu_2, \sigma_1^2, \sigma_2^2}_{\eta} \right)$$

*so that thee null hypothesis becomes $\psi = 0$.*

[1] This requires uniform bounds from Glivenko-Cantelli theorem and Donsker theorem that will be covered in STAT 582-583.

### 4.1.3 Wald test

The Wald test is perhaps the most intuitive approach to construct a test statistic. Essentially, it is just an application of the asymptotic normality of the MLE and normalizes the MLE into a $\chi^2$ test.

Let $\hat{\theta}_n = (\hat{\psi}_n, \hat{\eta}_n)$ be the MLE. Under conventional assumptions such as the QMD conditions, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$$

when $X_1, \cdots, X_n \sim P_\theta$.

Here you can see a nice feature for hypothesis test: under the null, we assume that the model is correct. So we can use the fact that Fisher's information matrix and the Hessian are inverse to each other:

$$I(\theta) \equiv \mathbb{E}_\theta \left( s(\theta|X_1) s(\theta|X_1)^T \right) = H^{-1}(\theta) \equiv \mathbb{E}_\theta \left( \nabla_\theta \nabla_\theta \log p(X_1; \theta) \right).$$

The asymptotic normality immediately implies the same thing for a subvector of $\hat{\theta}_n$, leading to

$$\sqrt{n}(\hat{\psi}_n - \psi) \xrightarrow{d} N(0, A^{-1}(\theta)),$$

where $A^{-1}(\theta)$ is the top $k \times k$ elements in $I^{-1}(\theta)$.

Under the null hypothesis of equation (4.4), $\psi = 0$ so we immediately have

$$\sqrt{n}\hat{\psi}_n \xrightarrow{d} N(0, A^{-1}(\theta)).$$

Since we can estimate $A(\theta)$ via $A(\hat{\theta}_n)$, Slutsky's theorem implies

$$\mathcal{W}_n \equiv n\hat{\psi}_n^T A(\hat{\theta}_n)\hat{\psi}_n \xrightarrow{d} \chi_k^2. \tag{4.5}$$

With this result, we reject $H_0$ if

$$\mathcal{W}_n > F_{\chi_k^2}^{-1}(1 - \alpha),$$

where $F_{\chi_k^2}$ is the cumulative distribution function of the $\chi_k^2$ distribution.

Note that we do not have to use $A(\hat{\theta}_n)$ as the covariance matrix estimator. Any other consistent estimator is sufficient in this case.

**Remark 4.2** *Suppose the Fisher's information matrix $I(\theta)$ takes the following block form:*

$$I(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix},$$

*where $I_{11}(\theta) \in \mathbb{R}^{k \times k}$. And assume that the inverse matrix*

$$I^{-1}(\theta) = \begin{pmatrix} A^{-1}(\theta) & M_{12}(\theta) \\ M_{21}(\theta) & C^{-1}(\theta) \end{pmatrix}.$$

*Then the simple matrix algebra shows that*

$$A(\theta) = I_{11}(\theta) - I_{12}(\theta) I_{22}^{-1}(\theta) I_{21}(\theta)$$

*as long as $k < d$. When $k = d$, we just write $A(\theta) = I(\theta)$.*

### 4.1.4 Likelihood ratio test

The likelihood ratio test (LRT) is another way to test the null hypothesis in equation (4.4). Let $L(\theta|X) = p(x;\theta)$ be the likelihood function and $\ell(\theta|x) = \log L(\theta|X)$ be the log-likelihood function and $s(\theta|x) = \nabla_\theta \ell(\theta|x)$ be the score function. For simplicity, we denote

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta|X_i)$$

and recall that the MLE $\hat{\theta}_n = \mathsf{argmax}_\theta \ell_n(\theta)$.

Formally, the LRT statistic is written as the following form:

$$\mathsf{LRT}_n = \frac{\sup_{\theta\in\Theta_0} \prod_{i=1}^n L(\theta|X_i)}{\sup_{\theta\in\Theta} \prod_{i=1}^n L(\theta|X_i)}.$$

Taking logarithm and multiplying it by $-2$ leads to

$$\mathcal{L}_n \equiv -2\log \mathsf{LRT}_n = 2n\ell_n(\hat{\theta}_n) - 2n \sup_{\theta\in\Theta_0} \ell_n(\theta) = 2n\ell_n(\hat{\theta}_n) - 2n\ell_n(\hat{\theta}_0),$$

where

$$\hat{\theta}_0 = \mathsf{argmax}_{\theta\in\Theta_0} \ell_n(\theta)$$

is the MLE under the null hypothesis. Note that by construction $\hat{\theta}_0 = (0, \hat{\eta}_0)$.

The new test statistic has the following nice Taylor expansion:

$$\begin{aligned}
\mathcal{L}_n &= 2n(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_0)) \\
&= -2n(\ell_n(\hat{\theta}_0) - \ell_n(\hat{\theta}_n)) \\
&= -2n(\hat{\theta}_0 - \hat{\theta}_n)^T \underbrace{\nabla_\theta \ell_n(\hat{\theta}_n)}_{=0} + n(\hat{\theta}_0 - \hat{\theta}_n)^T [\nabla_\theta \nabla_\theta \ell_n(\hat{\theta}_n)](\hat{\theta}_0 - \hat{\theta}_n) + o_P(1) \\
&= n(\hat{\theta}_0 - \hat{\theta}_n)^T I(\theta)(\hat{\theta}_0 - \hat{\theta}_n) + o_P(1) \\
&= \|\sqrt{n \cdot I(\theta)}(\hat{\theta}_0 - \hat{\theta}_n)\|^2 + o_P(1)
\end{aligned} \tag{4.6}$$

Therefore, the key to advance is to understand the vector $\sqrt{I(\theta)}(\hat{\theta}_0 - \hat{\theta}_n)$, the difference between the two MLEs (constrained versus unconstrained). First, recall from the M-estimation theory, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) = I^{-1}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\theta|X_i) + o_P(1) \tag{4.7}$$

for the unconstrained MLE $\hat{\theta}_n$.

Since $\hat{\theta}_0 = (0, \hat{\eta})^T$ is still the MLE but only on the parameter $\eta$, the same asymptotic theory applies, leading to

$$\sqrt{n}(\hat{\eta} - \eta) = I_{22}^{-1}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_2(\theta|X_i) + o_P(1), \tag{4.8}$$

where we decompose the score function into

$$s(\theta|x) = \begin{pmatrix} s_1(\theta|x) \\ s_2(\theta|x) \end{pmatrix},$$

with $s_1(\theta|x) \in \mathbb{R}^k$ and $s_2(\theta|x) \in \mathbb{R}^{d-k}$. The matrix $I_{22}^{-1}(\theta)$ is the inverse of $I_{22}(\theta)$, the top-left block matrix of $I(\theta)$, in Remark 4.2.

Combining equations (4.7) and (4.8), we have

$$\begin{aligned}\sqrt{n}I(\theta)(\hat{\theta}_0 - \hat{\theta}_n) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta)\end{pmatrix}\begin{pmatrix} 0 \\ I_{22}^{-1}(\theta)s_2(\theta|X_i)\end{pmatrix} - \begin{pmatrix} s_1(\theta|X_i) \\ s_2(\theta|X_i)\end{pmatrix}\right] + o_P(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} I_{11}(\theta)I_{22}^{-1}(\theta)s_2(\theta|X_i) - s_1(\theta|X_i) \\ 0\end{pmatrix} + o_P(1)\end{aligned}$$

This is a very powerful result–the lower part is asymptotically negligible compare to the top part (top $k$ components). Thus, we conclude that

$$\sqrt{n}I(\theta)(\hat{\theta}_0 - \hat{\theta}_n) \xrightarrow{d} \begin{pmatrix} Z(\theta) \\ 0\end{pmatrix}, \qquad Z \sim N(0, A(\theta)), \tag{4.9}$$

where $A(\theta)$ is the same top-right $k \times k$ matrix of $I^{-1}(\theta)$ as in the Wald test. This implies that the last quantity in equation (4.6) is

$$\|\sqrt{n \cdot I(\theta)}(\hat{\theta}_0 - \hat{\theta}_n)\|^2 = \|\sqrt{I^{-1}(\theta)}\sqrt{n}I(\theta)(\hat{\theta}_0 - \hat{\theta}_n)\|^2 \xrightarrow{d} Z(\theta)^T I^{-1}(\theta)Z(\theta) \overset{d}{=} \chi_k^2,$$

where for two random variables $U, V$, we write $U \overset{d}{=} V$ if they have identical distribution.

Putting this back to equation (4.6), we conclude that

$$\mathcal{L}_n \equiv -2\log \mathsf{LRT}_n = \|\sqrt{n \cdot I(\theta)}(\hat{\theta}_0 - \hat{\theta}_n)\|^2 + o_P(1) \xrightarrow{d} \chi_k^2. \tag{4.10}$$

Equation (4.10) is also known as the *Wilk's theorem*. With equation (4.10), we reject $H_0$ if $\mathcal{L}_n > F^{-1}_{\chi_k^2}(1-\alpha)$.

Informally, the Wilk's theorem says that:

> *If the null hypothesis puts $k$ equality constraints on the parameter space, then the test statistic $\mathcal{L}_n \equiv -2\log \mathsf{LRT}_n$ converges to $\chi_k^2$ under the null hypothesis.*

**Remark 4.3 (Geometry of the Wilk's theorem)** *Here is a more formal way to state the Wilk's theorem. If the true parameter $\theta \in \Theta_0$ such that there exists an open neighborhood around $\theta$ that $\Theta_0$ is a $(d-k)$-dimensional manifold, then the test statistic $\mathcal{L}_n \equiv -2\log \mathsf{LRT}_n \xrightarrow{d} \chi_k^2$. If $\Theta_0$ is formed by $k$ linearly separable equality constraints, then any interior point of $\Theta_0$ is locally a $(d-k)$-dimensional manifold by the implicit function theorem. This set $\Theta_0$ is known as a solution manifold.*

*Here is a geometric way to understand why the limiting distribution is a $\chi^2$ distribution with a degree of freedom $k$. Without the constraint, the MLE can go any direction, so after normalization, it behaves likes a Gaussian deviation from the true parameter $\theta$ in $d$ dimensions. Under the null hypothesis, the constrained MLE is a Gaussian vector on the $(d-k)$-dimensional manifold, which turns out to be the projection of the unconstrained MLE onto the manifold. The test statistic $\mathcal{L}_n$ is the length square between the two Gaussian vectors, which is asymptotically the projection of the unconstrained Gaussian vector onto the normal subspace of the manifold at $\theta$. Since the manifold has a dimension $(d-k)$, the normal direction forms a $k$-dimensional space, so the length square is a $\chi_k^2$ random variable.*

### 4.1.5 Score test

The score test is another population approach to form a test statistic. It literally uses the score equation to form the test statistic since under the null hypothesis, the score should converges to 0.

At a population level, the true parameter $\theta \in \Theta$ solves the score equation

$$\mathbb{E}_\theta(s(\theta|X)) = 0.$$

This motivates us to consider the normalized empirical score function

$$Z_n(\theta) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n s(\theta|X_i). \tag{4.11}$$

Coonsider

$$\mathcal{Z}_n \equiv Z_n(\hat{\theta}_0) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n s(\hat{\theta}_0|X_i). \tag{4.12}$$

One can easily see that the unconstrained MLE $\hat{\theta}_n$ solves the score equation $0 = Z_n(\hat{\theta}_n)$. So we now investigate the behavior of $\mathcal{Z}_n$. A simply Taylor expansion shows that

$$\begin{aligned}
\mathcal{Z}_n &= Z_n(\hat{\theta}_0) - Z_n(\hat{\theta}_n) \\
&= [\nabla Z_n(\hat{\theta}_n)](\hat{\theta}_0 - \hat{\theta}_n) + o_P(\|\hat{\theta}_0 - \hat{\theta}_n\|) \\
&= \left[\frac{1}{\sqrt{n}} \nabla Z_n(\hat{\theta}_n)\right] \sqrt{n}(\hat{\theta}_0 - \hat{\theta}_n) + o_P(\|\hat{\theta}_0 - \hat{\theta}_n\|).
\end{aligned}$$

Under the null hypothesis, $\hat{\theta}_0 \xrightarrow{P} \theta \in \Theta$, so we have

$$\frac{1}{\sqrt{n}} \nabla Z_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \nabla_\theta \nabla_\theta \ell(\hat{\theta}_n|X_i) \xrightarrow{P} \mathbb{E}_\theta(\nabla_\theta \nabla_\theta \ell(\theta|X_1)) \equiv H(\theta) = I(\theta).$$

Therefore, we can apply equation (4.9) again with Slutsky's theorem, which leads to

$$\mathcal{Z}_n \xrightarrow{d} N(0, A(\theta)).$$

Therefore, the score test statistic is

$$\mathcal{S}_n \equiv \mathcal{Z}_n^T I^{-1}(\hat{\theta}_0) \mathcal{Z}_n. \tag{4.13}$$

By Slutsky's theorem and continuous mapping theorem that $I^{-1}(\hat{\theta}_0) \xrightarrow{P} I^{-1}(\theta)$, we conclude that

$$\mathcal{S}_n \xrightarrow{d} \chi^2_k.$$

We can reject $H_0$ easily by the upper $1 - \alpha$ quantile of $\chi^2_k$.

### 4.1.6 Relation of the three test statistics

Now we have seen three popular test statistics from equations (4.5), (4.6), and (4.13):

$$\begin{aligned}
\mathcal{W}_n &\equiv n \hat{\psi}_n^T A(\hat{\theta}_n) \hat{\psi}_n, \\
\mathcal{L}_n &\equiv -2 \log \mathsf{LRT}_n = \frac{\sup_{\theta \in \Theta_0} \prod_{i=1}^n L(\theta|X_i)}{\sup_{\theta \in \Theta} \prod_{i=1}^n L(\theta|X_i)}, \\
\mathcal{S}_n &\equiv \mathcal{Z}_n^T I^{-1}(\hat{\theta}_0) \mathcal{Z}_n.
\end{aligned}$$

You can show that

$$\begin{aligned}\mathcal{W}_n - \mathcal{L}_n &\xrightarrow{P} 0,\\ \mathcal{L}_n - \mathcal{S}_n &\xrightarrow{P} 0,\\ \mathcal{S}_n - \mathcal{W}_n &\xrightarrow{P} 0,\end{aligned}$$

under $H_0$. So they are asymptotically equivalent when $H_0$ is true. However, in finite sample case, they have different performances.

## 4.2 Contiguity and Le Cam's lemmas

In the previous section, we have studied hypothesis tests under the null hypothesis. This allows us to control the type-1 errors. Now the next question we want to address is: if the null hypothesis is wrong, how do we know the performance of our test? Can we show that the power will eventually go to 1?

To answer this question, we need to study the behavior of a test statistic under the alternative hypothesis. As a case study, we consider a very simple hypothesis test problem.

**Example 4.4 (Normal simple versus simple)** *Suppose $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$, where $\sigma^2$ is known. We want to test a simple versus simple hypothesis:*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_0 + h$$

*for some fixed value $h$.*

*Let $P_\theta$ denote the model of $N(\theta, \sigma^2)$. The likelihood ratio for a single observation $X = x$ can be written as the following 'fancy' expression:*

$$\frac{dP_{\theta_0+h}}{dP_{\theta_0}}(x) = \frac{p_{\theta_0+h}(x)}{p_\theta(x)} = \exp\left(\frac{1}{2\sigma^2}\left[(x-\theta_0)^2 - (x-\theta_0-h)^2\right]\right) = \exp\left(\frac{2}{\sigma^2}[2h(x-\theta_0) - h^2]\right).$$

*Thus, the log-likelihood ratio over $X_1, \cdots, X_n$ is*

$$\Lambda_n = \log \prod_{i=1}^n \frac{dP_{\theta_0+h}}{dP_{\theta_0}}(X_i) = \frac{2n}{\sigma^2}[h(\bar{X}_n - \theta_0) - h^2] = \frac{nh(\bar{X}_n - \theta_0)}{\sigma^2} - \frac{nh^2}{2\sigma^2}.$$

*Under the null hypothesis, we have $\bar{X}_n \sim N(\theta_0, \sigma^2/n)$, which implies that*

$$\Lambda_n \sim N\left(-\frac{nh^2}{2\sigma^2}, \frac{nh^2}{\sigma^2}\right).$$

*Notice the interesting relation: the mean of $\Lambda_n$ is negative of half of the variance.*

The above example shows an interesting result of a likelihood ratio: we obtain an asymptotic log-normal distribution with mean equals to negative half of the variance. This interesting relation is not unique to this example–instead, it turns out to be an important result for several smooth models and it is a requirement for the famous Le Cam's third lemma (see Theorem 4.7). Before we proceed, we need to introduce a concept called *contiguity*.

### 4.2.1 Contiguity

For two probability measures $P$ and $Q$ on the same measurable space, we say $Q$ is *absolute continuous* with respect to $P$, denoted as $P \gg Q$ if for any measurable set $A$,

$$P(A) = 0 \Longrightarrow Q(A) = 0. \tag{4.14}$$

Simply put, if $P \gg Q$, you can think of the support of $P$ covers the support of $Q$. Therefore, the Radon-Nikodym derivative $\frac{dQ}{dP}$ is well-defined. We can easily define the 'likelihood ratio' $L(z) = \frac{dQ}{dP}(z)$ and

$$Q(A) = \int_A I(z \in A) L(z) dP(z). \tag{4.15}$$

The above formula means that if $P \gg Q$ and we know the likelihood ratio $L(z)$ and the distribution $P$, we are able to derive the probability model of $Q$. This is the change of measure formula.

However, there is one small problem. Our analysis of test statistics is asymptotic results, not finite sample results. Therefore, we need to generalize the concept of absolute continuity to asymptotic behavior.

For two sequences of probability measures $\{Q_n\}$ and $\{P_n\}$, we say $Q_n$ is *contiguous* with respect to $P_n$, denoted as $P_n \triangleright Q_n$, if for any sequence $\{A_n\}$,

$$P_n(A_n) \to 0 \Longrightarrow Q_n(A_n) \to 0.$$

We write $P_n \triangleleft\triangleright Q_n$ if $P_n \triangleright Q_n$ and $Q_n \triangleright P_n$; this is called mutual contiguity.

**Example 4.5** *Here are some interesting examples about contiguity and absolute continuity.*

- *Suppose* $P_n \overset{d}{=} N(0,1)$ *and* $Q_n \overset{d}{=} N(\mu_n, \sigma^2)$ *for some* $\sigma^2 > 0$. *If* $\mu_n \to \mu$, *then* $P_n \triangleleft\triangleright Q_n$.
- *Suppose* $P_n \overset{d}{=} N(0,1)$ *and* $Q_n \overset{d}{=} N(n, \sigma^2)$ *for some* $\sigma^2 > 0$. *Then* $P_n \gg Q_n$ *for every* $n$ *but we do NOT have* $P_n \triangleright Q_n$ *for the sequence* $A_n = [n, n+1]$.
- *Suppose* $P_n \overset{d}{=} \mathsf{Uni}[0,1]$ *and* $Q_n \overset{d}{=} \left[0 + \frac{(-1)^n}{n}, 1 + \frac{(-1)^n}{n}\right]$. *Then* $P_n \triangleleft\triangleright Q_n$ *but we do NOT have* $P_n \gg Q_n$ *nor* $Q_n \gg P_n$ *for any* $n$ *(since the interval is shifting).*

### 4.2.2 Le Cam's third lemma

Le Cam has studied the problem of contiguity throughly. To study the behavior of a test statistic under alternative, we only need to use a small lemma from him, known as Le Cam's third lemma.

Before we formally introduce this lemma, we first state an interesting result.

**Proposition 4.6 (Asymptotic log-normality; example 6.5 in van der Vaart)** *Consider two sequences of distributions* $P_n$ *and* $Q_n$ *such that the likelihood ratio*

$$L(X) = \frac{dQ_n}{dP_n}(X) \overset{d}{\to} e^{N(\mu, \sigma^2)}$$

*when* $X \sim P_n$, *then*

$$Q_n \triangleright P_n.$$

*Moreover,* $P_n \triangleleft\triangleright Q_n$ *if and only if* $\mu = -\frac{1}{2}\sigma^2$. *Note that* $X_n \overset{d}{\to} e^{N(\mu,\sigma^2)}$ *means that* $\log X_n \overset{d}{\to} N(\mu, \sigma^2)$.

While the log-normal with $\mu = -\frac{1}{2}\sigma^2$ may seem strange, it actually happens in several problems. One example is the simple versus simple normal model in Example 4.4.

**Theorem 4.7 (Le Cam's third lemma; example 6.7 of van der Vaart)** *Let $Z_n$ be a random vector from $P_n$ and assume that after transformation, the random vector $\Psi(Z_n)$ along with the likelihood ratio $L_n(z) = \frac{dQ_n}{dP_n}(z)$ satisfies*

$$\begin{pmatrix} \Psi(Z_n) \\ \log \frac{dQ_n}{dP_n}(Z_n) \end{pmatrix} \xrightarrow{d} N\left( \begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau & \sigma^2 \end{pmatrix} \right),$$

*then if $Z_n \sim Q_n$ the random vector*

$$\Psi(Z_n) \xrightarrow{d} N(\mu + \tau, \Sigma).$$

Theorem 4.7 is like an asymptotic version of the change of measure formula in equation (4.15). Notice that the log-likelihood ratio has an asymptotic normal distribution with mean $\mu = -\frac{1}{2}\sigma^2$, which is the implicit conclusion of Proposition 4.6.

The quantity $\Psi(Z_n)$ can be viewed as a random vector of interest. Its limiting distribution under $Q_n$ is the limiting distribution under $P_n$ plus a shift $\tau$, which is the covariance between $\Psi(Z_n)$ and the log-likelihood ratio under $P_n$. As you will see shortly, $\Psi(Z_n)$ will be our test statistic so all we need to do is to show the asymptotic normality of a test statistic along with the likelihood ratio statistic.

## 4.3 Local alternative

The local alternative model consider the following simple versus simple hypothesis problem:

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_0 + h_n$$

with $h_n \to 0$ at some rate. In reality, this rarely happens but studying this problem offers us insight into how a test will behave.

Clearly, the asymptotic analysis from the MLE tells us that when $h_n$ converges to 0 slower than $1/\sqrt{n}$, i.e., $\sqrt{n}h_n \to \infty$, the power $\pi_n(\theta)$ will go to 1 under $H_1$ since the separation between $H_0$ and $H_1$ is much higher than the MLE's error $O_P(1/\sqrt{n})$.

The interesting case to investigate is when $h_n$ is at exactly $1/\sqrt{n}$ rate. This is a famous scenario to look into. So we will consider the case $h_n = h/\sqrt{n}$.

The first thing we will apply is Theorem 3.14 (Theorem 7.2 of [van der Vaart]) of our previous lecture note: when $P_\theta$ is QMD at $\theta$, we have

$$\Lambda_n \equiv \log \prod_{i=1}^n \frac{dP_{\theta+h/\sqrt{n}}}{dP_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T s(\theta|X_i) - \frac{1}{2} h^T I(\theta) h + o_P(1) \tag{4.16}$$

when $X_1, \cdots, X_n \sim P_\theta$. This is a very significant result because the dominating random term in the right-hand-side is $\frac{1}{\sqrt{n}} \sum_{i=1}^n h^T s(\theta|X_i)$, which we clearly have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n h^T s(\theta|X_i) \xrightarrow{d} N\left(0, \underbrace{h^T \mathbb{E}(s(\theta|X_1)s(\theta|X_1)^T)h}_{=h^T I(\theta)h}\right).$$

This means that

$$\Lambda_n \xrightarrow{d} N\left(-\frac{1}{2}h^T I(\theta)h, h^T I(\theta)h\right),$$

which satisfies the requirement that mean is negative of half variance that is needed for Le Cam's third lemma (Theorem 4.7)!

Therefore, for any other statistic $T_n$, we just need to verify that jointly $(T_n, \Lambda_n)^T$ converges to a multivariate normal distribution and then compute the asymptotic covariance $\tau = \lim_{n\to\infty} \mathsf{Cov}(T_n, \Lambda_n)$ under the null hypothesis $P_\theta$. Then by Le Cam's third lemma (Theorem 4.7), under the $H_1$, the distribution of $T_n$ is simply the limiting normal distribution of $T_n$ under $H_0$ with the mean shifted by $\tau$.

Formally, this can be summarized as the following theorem.

**Theorem 4.8 (Local alternative)** *Consider the simple versus simple hypothesis tests:*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}.$$

*Assume $P_\theta$ is QMD at $\theta_0$ and the likelihood ratio $\Lambda_n$ in equation (4.16) satisfies*

$$\begin{pmatrix} Z_n \\ \Lambda_n \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mu_Z \\ -\frac{1}{2}h^T I(\theta)h \end{pmatrix}, \begin{pmatrix} \Omega & \tau \\ \tau^T & h^T I(\theta)h \end{pmatrix}\right).$$

*Then under the alternative hypothesis $\theta = \theta_0 + h/\sqrt{n}$,*

$$\Psi(Z_n) \xrightarrow{d} N(\mu_Z + \tau, \Omega).$$

While Theorem 4.8 may seem abstract, here is a simple application of it. We choose $Z_n$ to be the scaled MLE difference $Z_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$. Under the M-estimation theory (e.g., (Q1-Q4) in Theorem 3.10 of previous lecture), we immediately have

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \Lambda_n \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h^T I(\theta_0)h \end{pmatrix}, \begin{pmatrix} I^{-1}(\theta_0) & h \\ h^T & h^T I(\theta_0)h \end{pmatrix}\right).$$

The above result is just a simple combination of equation (4.7)

$$\sqrt{n}(\hat{\theta}_n - \theta) = I^{-1}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^n s(\theta|X_i) + o_P(1)$$

and equation (4.16). The covariance $h$ is obtained via the fact that

$$\begin{aligned}
\mathsf{Cov}(\sqrt{n}(\hat{\theta}_n - \theta_0), \Lambda_n) &\approx \mathbb{E}\left(\left[I^{-1}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^n s(\theta|X_i)\right]\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n h^T s(\theta|X_i)\right]\right) \\
&= \mathbb{E}\left(\left[I^{-1}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^n s(\theta|X_i)\right]\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n s^T(\theta|X_i)h\right]\right) \\
&= \mathbb{E}\left(I^{-1}(\theta)\left[\frac{1}{n}\sum_{i=1}^n s(\theta|X_i)s^T(\theta|X_i)\right]h\right) \\
&= h.
\end{aligned}$$

Therefore, we conclude that under the alternative hypothesis,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(h, I^{-1}(\theta_0)).$$

We will use this result to investigate the power of the Wald test.

**Remark 4.9 (Why not just do an asymptotic analysis on alternative hypothesis?)** *You may be wondering why we do not just perform the asymptotic analysis on the alternative hypothesis, i.e., assuming that the data is from $P_{\theta+h/\sqrt{n}}$ and work out the behavior of the MLE $\hat{\theta}_n$. Technically, you may do it this way but it will involve some non-trivial analysis. A major challenge we need to deal with is the fact that the distribution $P_{\theta+h/\sqrt{n}}$ that generates our data is changing with respect to $n$. While you can still apply Lindeberg-Feller's central limit theorem to achieve the asymptotic normality, a number of our conventional analysis will need some revision. Thus, using Le Cam's third lemma bypasses these complications and allow us to use the same condition, QMD, to derive the asymptotic normality.*

### 4.3.1 Power of Wald test

In this section, we formally analyze the power of the Wald test. Recall that the power is $\pi_n(\theta) = \mathbb{E}_\theta(\phi(X_1, \cdots, X_n))$ is the probability of rejecting null hypothesis when the data is generated from the model $p_\theta$ with a parameter $\theta$. Recall that we reparametrize the parameter so that $\theta = \begin{pmatrix} \psi \\ \eta \end{pmatrix}$ and $\psi \in \mathbb{R}^k$ and the null hypothesis is $H_0 : \psi = 0$.

In Section 4.1.3, we have shown that we can control the type-1 error by rejecting null when

$$\mathcal{W}_n \equiv n\hat{\psi} A(\hat{\theta})\hat{\psi} > F^{-1}_{\chi^2_k}(1-\alpha),$$

where $\hat{\psi}$ is the top $k$ element of the MLE.

**Theorem 4.10 (Power of Wald test)** *Let $\theta_0 \in \Theta_0$. Assume the following:*

- *$P_\theta$ is QMD at $\theta_0$,*
- *$I(\theta)$ is non-singular and continuous at $\theta = \theta_0$,*
- *$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} I^{-1}(\theta_0) \sum_{i=1}^n s(\theta_0|X_i) + o_P(1)$ when the data is from $P_{\theta_0}$.*

*Then under the local alternative $\theta = \theta_0 + h/\sqrt{n}$, we have*

$$\mathcal{W}_n \equiv n\hat{\psi} A(\hat{\theta})\hat{\psi} \xrightarrow{d} \chi^2_k(h_k^T A(\theta_0) h_k),$$

*where $h_k$ is the top $k$ element of $h$ and $\chi^2_k(c)$ is the non-central $\chi^2$ distribution with a degree of freedom $k$ and non-centrality $c$.*

To be more specific, the non-central $\chi^2_k(c)$ distribution is defined as follows. Recall that a $\chi^2_k$ random variable $W_k$ can be expressed as

$$W_k \overset{d}{=} Z_1^2 + \cdots + Z_k^2,$$

where $Z_1, \cdots, Z_k \sim N(0,1)$. The non-central $\chi^2_k(c)$ is the similar random variable but each $Z_j \sim N(a_j, 1)$ and $c = \sum_{j=1}^k a_j^2$.

Since we reject the null when $\mathcal{W}_n > F^{-1}_{\chi^2_k}(1-\alpha)$, i.e.,

$$\phi(X_1, \cdots, X_n) = I\left(\mathcal{W}_n > F^{-1}_{\chi^2_k}(1-\alpha)\right),$$

we immediately have

$$\pi_n(\theta) = P\left(\chi^2_k > F^{-1}_{\chi^2_k}(1-\alpha))\right) = P\left(\chi^2_k(0) > F^{-1}_{\chi^2_k}(1-\alpha))\right) = \alpha$$

for any $\theta \in \Theta_0$. When $\theta = \theta_0 + h/\sqrt{n} \notin \Theta_0$, we have

$$\lim_{n\to\infty} \pi_n(\theta) = P\left(\chi^2_k(h_k^T A(\theta_0) h_k) > F^{-1}_{\chi^2_k}(1-\alpha))\right) > P\left(\chi^2_k(0) > F^{-1}_{\chi^2_k}(1-\alpha))\right) = \alpha.$$

## 4.4 Relative efficiency

Our analysis of the test power relies on Theorem 4.8, or more generally, the Le Cam's third lemma. But this theorem is not limited to the MLE, it works for any estimator that is asymptotically normal.

Now we consider a generic quantity $\mu(\theta) \in \mathbb{R}^m$. Let $\hat{\mu}_n$ be an asymptotic linear estimator of $\mu(\theta)$ with an influence function $g_\theta$, i.e.,

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_\theta(X_i) + o_P(1).$$

Under the null hypothesis, i.e., data $X_1, \cdots, X_n \sim P_\theta$ with $\theta \in \Theta_0$, we assume that

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_\theta(X_i) + o_P(1) \xrightarrow{d} N(0, \mathbb{E}_\theta(g_\theta(X_1) g_\theta(X_1)^T)). \tag{4.17}$$

Equation (4.17) is generally easy to obtain since we only need to focus on cases where the null hypothesis is correct. Now we want to investigate the normality under the local alternative:

$$\theta = \theta_0 + h/\sqrt{n},$$

where $\theta_0 \in \Theta_0$.

**Theorem 4.11** *Assume that $P_\theta$ is QMD at $\theta_0$ and $\hat{\mu}_n$ satisfies equation (4.17) when $\theta = \theta_0$. For a given $h \in \mathbb{R}^d$, we have the following :*

$$\begin{pmatrix} \sqrt{n}(\hat{\mu}_n - \mu(\theta)) \\ \Lambda_n \end{pmatrix} \xrightarrow{d} N\left( \begin{pmatrix} 0 \\ -\frac{1}{2} h^T I(\theta) h \end{pmatrix}, \begin{pmatrix} \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1) g_{\theta_0}(X_1)^T) & \tau_h \\ \tau_h^T & h^T I(\theta) h \end{pmatrix} \right),$$

*where $\tau_h = \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1) s(\theta_0 | X_1)^T h)$. Thus, for the sequence $\theta = \theta_0 + h/\sqrt{n}$, we have*

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta_0)) \xrightarrow{d} N\left(\tau_h, \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1) g_{\theta_0}(X_1)^T)\right).$$

Theorem 4.11 shows a powerful result–even if we only show asymptotic normality at $\theta = \theta_0$, we can generalize it into a neighborhood of it $\theta = \theta_0 + h/\sqrt{n}$. Notice that there is a drift therm $\tau_h$, which seems to be caused by the fact that we are using $\hat{\mu}_n$ to estimate $\mu(\theta_0)$ while the truth is $\theta = \theta_0 + h/\sqrt{n}$.

Now suppose we want to estimate the population characteristic $\mu(\theta)$, not $\mu(\theta_0)$. Theorem 4.11 implies that

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta_0 + h/\sqrt{n})) \xrightarrow{d} N\left(\tau_h + \sqrt{n}\left[\mu(\theta_0 + h/\sqrt{n}) - \mu(\theta_0)\right], \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1)g_{\theta_0}(X_1)^T)\right).$$

Now suppose $\mu(\theta)$ is differentiable at $\theta_0$, we then have

$$\sqrt{n}\left[\mu(\theta_0 + h/\sqrt{n}) - \mu(\theta_0)\right] = h^T\nabla\mu(\theta_0) + o(1).$$

Therefore, if we have the following condition

$$\tau_h - h^T\nabla\mu(\theta_0) \equiv \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1)s(\theta_0|X_1)^T h) - h^T\nabla\mu(\theta_0) = 0,$$

or equivalently,

$$\mathbb{E}_{\theta_0}(g_{\theta_0}(X_1)s(\theta_0|X_1)^T) = \nabla\mu(\theta_0), \tag{4.18}$$

we have

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta_0 + h/\sqrt{n})) \xrightarrow{d} N(0, \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1)g_{\theta_0}(X_1)^T)),$$

where the limiting distribution does NOT depend on $h$.

**Regular estimator.** The fact that the limiting distribution of an estimator does not depend on the direction $h$ that is approaching the limit is called a *regular estimator*. Formally, the estimator $\hat{\mu}_n$ is *regular* at $P_\theta$ with $\theta = \theta_0$ if for any $h \in \mathbb{R}^d$,

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta_0 + h/\sqrt{n})) \xrightarrow{d} Z$$

when data is from $P_{\theta_0 + h/\sqrt{n}}$ and $Z$ does not depend on $h$.

In the above analysis, we see that a sufficient condition for $\hat{\mu}_n$ to be a regular estimator is that its influence function $g_\theta$ must satisfies equation (4.18).

### 4.4.1 Power of simple regular estimator

The introduction of the transformation $\mu$ gives us a simple way to compare powers of regular estimators.

Formally, when $m = 1$, we have

$$\sqrt{n}(\hat{\mu}_n - \mu(\theta_0)) \xrightarrow{d} N\left(0, \sigma_g^2(\theta_0)\right),$$

where $\sigma_g^2(\theta_0) = \mathbb{E}_{\theta_0}(g_{\theta_0}^2(X_1))$.

To simplify the notations, we consider a one-sided test and a simple null that $H_0 : \theta = \theta_0$. For simplicity, we assume $\mu(\theta_0) = 0$ so we reject the null hypothesis if

$$\frac{\sqrt{n}\hat{\mu}_n}{\sigma_g(\theta_0)} > z_{1-\alpha}, \tag{4.19}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of $N(0,1)$. Namely, our test function is

$$\phi(X_1, \cdots, X_n) = I\left(\frac{\sqrt{n}\hat{\mu}_n}{\sigma_g(\theta_0)} > z_{1-\alpha}\right).$$

Clearly, under $H_0$, $\pi_n(\theta_0) = \mathbb{E}_{\theta_0}(\phi(X_1, \cdots, X_n)) = \alpha$ so we control the type-1 error.

Now we consider $\theta = \theta_0 + h/\sqrt{n}$ and investigate the power.

Assume that equation (4.18) holds so the estimator $\hat{\mu}_n$ is regular. Then we immediately have

$$\sqrt{n}(\hat{\mu}_n - \sqrt{n}\mu(\theta_0 + h/\sqrt{n})) \xrightarrow{d} N\left(0, \sigma_g^2(\theta_0)\right)$$

when data is from $P_{\theta_0+h/\sqrt{n}}$. Thus,

$$\sqrt{n}\hat{\mu}_n \approx N\left(\sqrt{n}\mu(\theta_0 + h/\sqrt{n}), \sigma_g^2(\theta_0)\right)$$

and using the fact that $\mu(\theta_0) = 0$, we have

$$\sqrt{n}\mu(\theta_0 + h/\sqrt{n}) = \sqrt{n}[\mu(\theta_0 + h/\sqrt{n}) - \mu(\theta_0)] = h^T\nabla\mu(\theta_0) + o(1).$$

Therefore, we have

$$\sqrt{n}\hat{\mu}_n \approx N\left(h^T\nabla\mu(\theta_0), \sigma_g^2(\theta_0)\right),$$

which is a shifted Gaussian distribution. Note that we use the approximation notation $\approx$ in the above derivation. All these derivations can be done more rigorously using convergence in distribution and probability.

Thus, the power at $\theta = \theta_0 + h/\sqrt{n}$ is

$$\begin{aligned}\pi_n(\theta_0 + h/\sqrt{n}) &= P\left(\frac{\sqrt{n}\hat{\mu}_n}{\sigma_g(\theta_0)} > z_{1-\alpha}\right) \\ &= P\left(\frac{\sqrt{n}\hat{\mu}_n - h^T\nabla\mu(\theta_0)}{\sigma_g(\theta_0)} > z_{1-\alpha} - \frac{h^T\nabla\mu(\theta_0)}{\sigma_g(\theta_0)}\right) \\ &\to 1 - \Phi\left(z_{1-\alpha} - \frac{h^T\nabla\mu(\theta_0)}{\sigma_g(\theta_0)}\right),\end{aligned} \tag{4.20}$$

where $\Phi(t)$ is the CDF of $N(0,1)$ such that $\Phi(z_\beta) = \beta$.

Here you can see that if $h$ is in the direction of $\nabla\mu(\theta_0)$, i.e., $h^T\nabla\mu(\theta_0) > 0$, we will obtain

$$z_{1-\alpha} - \frac{h^T\nabla\mu(\theta_0)}{\sigma_g(\theta_0)} < z_{1-\alpha}$$

so $\pi_n(\theta_0 + h/\sqrt{n}) > 1 - \Phi(z_{1-\alpha}) = \alpha$ so we gain power (relative to the power at null $\alpha$). On the other hand, if $h^T\nabla\mu(\theta_0) < 0$, we end up losing power. Note that this is caused by the fact that we are using a one-sided test. If we consider a two-sided test, we will always have a higher power than $\alpha$.

**Example 4.12 (Location family: sign test versus t-test)** *Our data $X_1, \cdots, X_n \sim P_\theta$ where $P_\theta$ has a PDF $f(x-\theta)$ such that $f$ is a known function and $X_i \in \mathbb{R}$. Assume the following conditions:*

- $v^2 = \int x^2 f(x)dx < \infty$.
- *$f$ is symmetric at 0, so $\theta$ is both the mean and median of $P_\theta$.*
- *$f$ is positive and continuously differentiable and $\int \frac{|f'(x)|^2}{f(x)}dx < \infty$*[2].

*We consider the following hypothesis:*

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0.$$

*We now compare the sign test and the t-test. The sign-test uses the test statistic*

$$S_n = \frac{1}{n}\sum_{i=1}^n I(X_i > 0)$$

[2]This implies that $P_\theta$ is QMD; see Example 7.8 of [van der Vaart].

*while the t-test uses*

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\hat{\sigma}_n},$$

*where* $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

*You can easily show that under the null,*

$$\begin{aligned}\sqrt{n}\left(S_n - \frac{1}{2}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_0(X_i) + o_P(1) \xrightarrow{d} N(0, 1/4),\\ \sqrt{n}T_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} t_0(X_i) + o_P(1) \xrightarrow{d} N(0,1),\end{aligned} \tag{4.21}$$

*where the influence functions are* $g_S(x) = I(x > 0) - \frac{1}{2}$ *and* $g_T(x) = \frac{x}{v}$.

*To apply the above analysis we have done, we need to show that both estimator* $S_n$ *and* $T_n$ *are regular with respect to their parameters of interest. The result in equation (4.21) shows that* $\mu(0) = \frac{1}{2}$ *for* $S_n$ *while* $\mu(0) = 0$ *for* $T_n$.

*Now we need to show equation (4.18), i.e.,*

$$\mu'(\theta_0) = \mathbb{E}_{\theta_0}(g_{\theta_0}(X_1)s(\theta_0|X_1))$$

*with* $\theta_0 = 0$.

*For the sign-test, the parameter of interest is clearly* $\mu(\theta) = P_\theta(X > 0)$. *Thus,*

$$\begin{aligned}\mu'(0) &= \frac{d}{d\theta}P_\theta(X > 0)\Big|_{\theta=0}\\ &= \frac{d}{d\theta}\int_0^\infty f(x-\theta)dx\Big|_{\theta=0}\\ &= \int_0^\infty \frac{\partial}{\partial\theta}f(x-\theta)\Big|_{\theta=0}dx\\ &= -\int_0^\infty f'(x)dx.\end{aligned}$$

*Now the score function* $s(\theta|x) = \frac{\partial}{\partial\theta}\log f(x-\theta) = -\frac{f'(x-\theta)}{f(x-\theta)}$ *so* $s(0|x) = -\frac{f'(x)}{f(x)}$ *and* $g_{\theta=0}(x) = g_S(x) = I(x > 0)$. *This implies*

$$\begin{aligned}\mu'(0) &= -\int_0^\infty f'(x)dx\\ &= \int I(x>0) - f'(x)dx\\ &= \int I(x>0) - \frac{f'(x)}{f(x)}f(x)dx\\ &\mathbb{E}_{\theta_0}(g_S(X_1)s(\theta_0|X_1)).\end{aligned}$$

*So the sign-test statistic* $S_n$ *is regular for estimating* $\mu(\theta)$ *at* $\theta = 0$.

*You can show that the t-test is also regular at its own* $\mu(0)$ *with* $\mu'(0) = 1/v = 1/\sqrt{\int x^2 f(x)dx}$.

*Now we compare their power using equation (4.20). For sign-test, $\sigma_g(0) = \frac{1}{2}$ from its limiting distribution and $\mu'(0) = -\int_0^\infty f'(x)dx = f(0)$. Thus, its power under $\theta = 0 + h/\sqrt{n}$ is*

$$\pi_n(h/\sqrt{n}) = P_{\theta=h/\sqrt{n}}(2 \times \sqrt{n}\left(S_n - \frac{1}{2}\right) > z_{1-\alpha}) \to 1 - \Phi(z_{1-\alpha} - 2hf(0)).$$

*For t-test, $\sigma_g(0) = 1$ and $\mu'(0) = 1/v$, so its power under $\theta = 0 + h/\sqrt{n}$ is*

$$\pi_n(h/\sqrt{n}) = P_{\theta=h/\sqrt{n}}(\sqrt{n}T_n > z_{1-\alpha}) \to 1 - \Phi(z_{1-\alpha} - h/v).$$

*Note that $v^2 = \int x^2 f(x)dx$ is the variance under the null so $v$ is the standard deviation under the null.*

*Thus, the power between sign-test versus t-test depends on how $2f(0)$ versus $1/v$ with $v = \sqrt{\int x^2 f(x)dx}$. If $2f(0)v > 1$, then sign-test is better. If $2f(0)v < 1$, then the t-test is better.*

*Here are some interesting examples:*

- *Laplace family: $f_a(x) = e^{-|x|/a}$ for some $a > 0$. Then $2f(0)v = 2$ so sign-test is better.*
- *Normal family: $f_a(x) = \frac{1}{\sqrt{2\pi a^2}} e^{-\frac{1}{2a^2}x^2}$ for some $a > 0$. Then $2f(0)v \approx 0.82$ so t-test is better.*
- *Uniform family: $f_a(x) = \frac{1}{2a} I\left(\frac{|x|}{a} \le 1\right)$ for some $a > 0$. Then $2f(0)v \approx 0.33$ so t-test is much better.*

*You can see the trend is what we expect: if the tail is heavy (Laplace or even Cauchy), the sign-test is better than t-test. On the other hand, if the tail is light, the t-test is generally better.*

# A Information criteria and model selection

Suppose that we observe $X_1, \cdots, X_n$ from an unknown distribution function. The model selection problem occurs when we have multiple models for the data generating distribution. Here is one example.

**Example 4.13** *Suppose we observe $X_1, \cdots, X_n$ and we see that more observations concentrate around 0. We have a number of possible models for the underlying distribution:*

- *$\mathcal{M}_1$ : we assume that the data is from $N(0, \sigma^2)$ with unknown $\sigma^2$.*
- *$\mathcal{M}_2$ : we assume that the data is from $N(\mu, \sigma^2)$ with unknown $\mu, \sigma^2$.*
- *$\mathcal{M}_3$ : we assume that the data is from a double exponential with rate parameter $\lambda$, i.e., $p_\lambda(x) = \frac{\lambda}{2} e^{\lambda|x|}$.*
- *$\mathcal{M}_4$ : we assume that the data is from a double exponential with rate parameter $\lambda$ and center $\mu$, i.e., $p_{\lambda,\mu}(x) = \frac{\lambda}{2} e^{\lambda|x-\mu|}$.*

*In this case, model $\mathcal{M}_1$ is nested in $\mathcal{M}_2$, i.e., $\mathcal{M}_1 \subset \mathcal{M}_2$ and similarly, $\mathcal{M}_3 \subset \mathcal{M}_4$.*

*Model selection problem: When we are given the data, how can we choose the model that best fit the data?*

For a given model $p_\theta$, we denote

$$\ell_n = \ell(\hat{\theta}_n | X_1, \cdots, X_n) = \sum_{i=1}^n \ell(\hat{\theta}_n | X_i)$$

as the maximal value of the empirical log-likelihood function. Note that $\ell_n$ differs from models to models. In the example 4.13, we have four distinct values of $\ell_n$.

Intuitively, we may want to choose the model with highest value $\ell_n$. However, this choice suffers from *overfitting*: Suppose $\mathcal{M}_1$ is the correct model that the data is indeed from a mean 0 Gaussian, the maximal likelihood value $\ell_n$ at $\mathcal{M}_2$ will be higher than $\mathcal{M}_1$ since we have more parameters to optimize!

To avoid overfitting, a natural approach is to introduce a penalization term such that we will favor a simpler model. Namely, we will be using something like

$$R_k = -2\ell_n + P(\mathcal{M}_k)$$

such that $P(\mathcal{M}_k)$ increases with respect to the model complexity of $\mathcal{M}_k$ and we simply choose the model $k^*$ by minimizing $R_k$. Note that we use minus of the likelihood value so that $R_k$ behaves like a risk function that we want to minimize. The multiplication of 2, as you may have expected from our previous analysis on the log-likelihood function, will be related to the second-order Taylor expansion.

This seems to be a reasonable approach but the real question is:

> How should we define the penalty on the model complexity $P(\mathcal{M}_k)$?

## A.1 AIC: Akaike information criterion

The AIC is an information criterion that is common used for model selection. The AIC propose the following criterion:

$$AIC(\mathcal{M}) = 2d - 2\ell_n,$$

where $d$ is the dimension of the model $\mathcal{M}$, i.e., number of parameters. Namely, AIC chooses the penalty $P(\mathcal{M}_k) = d_k$ to be the number of parameter of model $\mathcal{M}_k$.

The idea of AIC is to adjust the empirical risk to be an unbiased estimator of the true risk in a parametric model. Under a likelihood framework, the loss function is the negative log-likelihood function so the empirical risk is

$$\hat{R}_n(\hat{\theta}_n) = -\ell_n = -\ell(\hat{\theta}_n | X_1, \cdots, X_n) = -\ell_n(\hat{\theta}_n).$$

On the other hand, the true risk of the MLE is

$$R(\hat{\theta}_n) = \mathbb{E}(-n\bar{\ell}(\hat{\theta}_n)).$$

Note that we multiply it by $n$ to reflect the fact that in the empirical risk, we did not divide it by $n$.

To derive the AIC, we examine the asymptotic bias of the empirical risk $\hat{R}_n(\hat{\theta}_n)$ versus the true risk $R(\hat{\theta}_n)$.

**Analysis of true risk.** To analyze the true risk $R(\hat{\theta}_n)$, we first investigate the asymptotic behavior of $\bar{\ell}(\hat{\theta}_n)$ around $\theta^*$:

$$\begin{aligned}\bar{\ell}(\hat{\theta}_n) &\approx \bar{\ell}(\theta^*) + (\hat{\theta}_n - \theta^*)^T \underbrace{\nabla\bar{\ell}(\theta^*)}_{=0} + \frac{1}{2}(\hat{\theta}_n - \theta^*)^T \underbrace{\nabla\nabla\bar{\ell}(\theta^*)}_{=I(\theta^*)}(\hat{\theta}_n - \theta^*) \\ &= \bar{\ell}(\theta^*) + \frac{1}{2}(\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*).\end{aligned}$$

Thus, the true risk is

$$R(\hat{\theta}_n) = -n\mathbb{E}(\bar{\ell}(\hat{\theta}_n)) \approx -n\bar{\ell}(\theta^*) - \frac{n}{2}\mathbb{E}\left((\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*)\right). \tag{4.22}$$

**Analysis of expected empirical risk.** For the expected empirical risk, we first expand $\ell_n$ as follows:

$$\begin{aligned}\ell_n &= \sum_{i=1}^n \ell(\hat{\theta}_n|X_i) \\ &\approx \sum_{i=1}^n \ell(\theta^*|X_i) + \underbrace{(\hat{\theta}_n - \theta^*)^T \sum_{i=1}^n \nabla\ell(\theta^*|X_i)}_{(I)} + \underbrace{\frac{1}{2}(\hat{\theta}_n - \theta^*)^T \sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i)(\hat{\theta}_n - \theta^*)}_{=(II)}.\end{aligned} \tag{4.23}$$

The expectation of the first quantity is $\bar{\ell}(\theta^*)$, which agrees with the first term in the true risk so all we need is to understand the behavior of the rest two quantities.

For the first quantity, using the fact that $\sum_{i=1}^n \nabla\ell(\hat{\theta}_n|X_i) = 0$,

$$\begin{aligned}\sum_{i=1}^n \nabla\ell(\theta^*|X_i) &= \sum_{i=1}^n \nabla(\ell(\theta^*|X_i) - \ell(\hat{\theta}_n|X_i)) \\ &\approx \left(\sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i)\right)(\theta^* - \hat{\theta}_n) \\ &\approx \underbrace{(\nabla\nabla\mathbb{E}(\ell(\theta^*|X_i)))}_{=I(\theta^*)} n(\theta^* - \hat{\theta}_n).\end{aligned}$$

Thus,

$$(I) \approx -n(\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*).$$

For quantity (II), note that

$$\frac{1}{n}\sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i) \approx \nabla\nabla\mathbb{E}(\ell(\theta^*|X_i)) = I(\theta^*)$$

so

$$\begin{aligned}(II) &= \frac{1}{2}(\hat{\theta}_n - \theta^*)^T \sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i)(\hat{\theta}_n - \theta^*) \\ &= \frac{n}{2}(\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*)\end{aligned}$$

Combining (I) and (II) into equation (4.23) and taking the expectation, we obtain

$$\mathbb{E}(\hat{R}_n(\hat{\theta}_n)) = -\mathbb{E}(\ell_n) = -n\bar{\ell}(\theta^*) + \frac{n}{2}\mathbb{E}\left((\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*)\right)$$

Comparing this to equation (4.22), we obtain

$$\mathbb{E}(\hat{R}_n(\hat{\theta}_n)) - R(\hat{\theta}_n) = -n\mathbb{E}\left((\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*)\right).$$

Note that by the theory of MLE, one can easily shown that

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \approx N(0, I^{-1}(\theta^*))$$

so

$$n(\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*) \approx \chi^2_d,$$

which implies that[3]

$$n\mathbb{E}\left((\hat{\theta}_n - \theta^*)^T I(\theta^*)(\hat{\theta}_n - \theta^*)\right) = d.$$

Thus, to make sure that we have an asymptotic unbiased estimator of the true risk $R(\hat{\theta}_n)$, we need to modify the empirical risk by

$$\hat{R}_n(\hat{\theta}_n) + d = -\ell_n + d.$$

Multiplying this quantity by 2, we obtain the AIC

$$AIC = 2d - 2\ell_n.$$

From the derivation of AIC, we see that the goal of the AIC is to adjust the model so that we are comparing unbiased estimates of the true risks across different models. Thus, the model selected by minimizing the AIC can be viewed as the model selected by minimizing unbiased estimates of the true risks. From the risk minimization point of view, this is trying to make a prediction using a good risk estimator. Thus, some people would common that the design of AIC is to choose a model that makes good predictions.

## A.2 BIC: Bayesian information criterion

Another common approach for model selection is the BIC:

$$BIC = d\log n - 2\ell_n,$$

where again $d$ denotes the dimension of the model. Namely, BIC chooses the penalty $P(\mathcal{M}_k) = d_k \log n$, so it penalizes a little more than the AIC.

Here is the derivation of the BIC. In the Bayesian setting, we place a prior $\pi(m)$ over all possible models and within each model, we place a prior over parameters $p(\theta|m)$. The BIC is a Bayesian criterion, which means that we will select model according to the posterior distribution of each model $m$. Namely, we will try to derive $\pi(m|X_1, \cdots, X_n)$.

By Bayes rule, we have

$$\pi(m|X_1, \cdots, X_n) = \frac{\pi(m, X_1, \cdots, X_n)}{p(X_1, \cdots, X_n)} \propto p(X_1, \cdots, X_n|m)\pi(m)$$

so all we need is to derive the marginal density in a model $p(X_1, \cdots, X_n|m)$.

With a prior $\pi(\theta|m)$, this marginal density can be written as

$$p(X_1, \cdots, X_n|m) = \int p(X_1, \cdots, X_n|\theta, m)\pi(\theta|m)d\theta. \tag{4.24}$$

Suppose that the model $m$ is correct in the sense that under model $m$, there exists $\theta^* \in \Theta$ such that the data are indeed generated from $p(x|\theta^*)$.

Using the log-likelihood function, we can expand

$$p(X_1, \cdots, X_n|\theta, m) = e^{\ell(\theta|X_1, \cdots, X_n, m)} = e^{\sum_{i=1}^n \ell(\theta|X_i, m)}. \tag{4.25}$$

[3]Formally, we need a few more conditions; the convergence in distribution is not enough for the convergence in expectation.

Asymptotically, the log-likelihood function can be further expand

$$\begin{aligned}\sum_{i=1}^n \ell(\theta|X_i,m) &= \sum_{i=1}^n \ell(\theta^*|X_i,m) + (\theta-\theta^*)^T \underbrace{\sum_{i=1}^n \nabla \ell(\theta^*|X_i,m)}_{=0} \\ &\quad + (\theta-\theta^*)^T \sum_{i=1}^n \nabla\nabla \ell(\theta^*|X_i,m)(\theta-\theta^*) + \text{ small terms} \\ &= \ell_n + n(\theta-\theta^*)^T \underbrace{\frac{1}{n}\sum_{i=1}^n \nabla\nabla \ell(\theta^*|X_i,m)}_{\approx -I(\theta^*)}(\theta-\theta^*) + \text{ small terms},\end{aligned}$$

where $I(\theta^*)$ is the Fisher's information matrix. Plugging this into equation (4.25) and ignoring the reminder terms, we obtain

$$p(X_1,\cdots,X_n|\theta,m) \approx e^{\ell_n - n(\theta-\theta^*)^T I(\theta^*)(\theta-\theta^*)}. \tag{4.26}$$

Thus, equation (4.24) can be rewritten as

$$p(X_1,\cdots,X_n|m) \approx e^{\ell_n} \int e^{-n(\theta-\theta^*)^T I(\theta^*)(\theta-\theta^*)} \pi(\theta|m) d\theta. \tag{4.27}$$

To compute the above integral, consider a random vector $Y \sim N(0, \frac{1}{n} I(\theta^*))$. The expectation

$$\begin{aligned}\mathbb{E}(\pi(Y|m)) &= \left(\frac{n}{2\pi}\right)^{d/2} \mathsf{det}^{-1}(I(\theta^*)) \int e^{-n(y-\theta^*)^T I(\theta^*)(y-\theta^*)} \pi(y|m) dy \\ &\approx \pi(\hat{\theta}_n|m)\end{aligned}$$

when $n \to \infty$. This implies that

$$\int e^{-n(\theta-\theta^*)^T I(\theta^*)(\theta-\theta^*)} \pi(\theta|m) d\theta \approx \left(\frac{2\pi}{n}\right)^{d/2} \mathsf{det}(I(\theta^*)) \pi(\hat{\theta}_n|m).$$

Putting this into equation (4.27), we conclude that the Bayesian evidence

$$p(X_1,\cdots,X_n|m) \approx e^{\ell_n} \left(\frac{2\pi}{n}\right)^{d/2} \mathsf{det}(I(\theta^*)) \pi(\hat{\theta}_n|m)$$

so the log evidence is

$$\log p(X_1,\cdots,X_n|m) \approx \ell_n - \frac{d}{2}\log n + \frac{d}{2}\log(2\pi) + \log \mathsf{det}(I(\theta^*)) + \log \pi(\hat{\theta}_n|m).$$

The only quantity that would increases with respect to the sample size $n$ are the first two quantities so after multiplying by $-2$ and keeping only the dominating two terms, we obtain

$$BIC = d \log n - 2\ell_n.$$

Although the BIC leads to a criterion similar to the AIC, the reasoning is somewhat different. In the construction of BIC, the effect of priors are ignored since we are working on the limiting regime but we still use the Bayesian evidence as a model selection criterion. We are selecting the model with the highest evidence. When the data is indeed generated from one of the model in the collection of models we are choosing from, the posterior will concentrate on this correct model. So BIC would eventually be able to select this model. Therefore, some people would argue that unlike AIC that chooses the best predictive model, the BIC attempts to select the true model if it exists in the model set.

## A.3 Model selection consistency

Now we prove that the BIC selects the correct model with a probability tending to 1 under the nested condition. It turns out that the additional $\log n$ factor in the BIC really helps in choosing the correct model.

Since we will be working with different models $\mathcal{M}_1, \cdots, \mathcal{M}_K$, and each of them has its own set of parameters, we have to be careful about the notations.

Each model

$$\mathcal{M}_k = \left\{ p_{\theta_{[k]}} : \theta_{[k]} \in \Theta_{[k]} \subset \mathbb{R}^{d_k} \right\},$$

where $d_k$ is the number of parameters under model $\mathcal{M}_k$.

The maximal log-likelihood value of model $\mathcal{M}_k$ is

$$\ell_{n,k} = \sum_{i=1}^{n} \log p_{\hat{\theta}_{[k]}}(X_i) = \sum_{i=1}^{n} \ell_{[k]}(\theta_{[k]}|X_i),$$

where $\hat{\theta}_{[k]}$ is the MLE under model $\mathcal{M}_k$

$$\hat{\theta}_{[k]} = \mathsf{argmax}_{\theta_{[k]}} \quad \sum_{i=1}^{n} \log p_\theta(X_i) = \mathsf{argmax}_{\theta_{[k]}} \sum_{i=1}^{n} \ell_{[k]}(\theta_{[k]}|X_i)$$

and we use $\ell_{[k]}(\theta_{[k]}|X_i)$ to denote the log-likelihood of parameter $\theta_{[k]}$ under model $\mathcal{M}_k$.

Formally, the BIC selects the model $\mathcal{M}_{\hat{k}}$

$$\hat{k} = \mathsf{argmax}_{k=1,\cdots,K} \quad \underbrace{d_k \log n - 2\ell_{n,k}}_{=BIC_{n,k}}. \tag{4.28}$$

For model $\mathcal{M}_k$, let

$$\theta^*_{[k]} = \mathsf{argmax}_{\theta_{[k]}} \quad \bar{\ell}_{[k]}(\theta_{[k]}), \qquad \bar{\ell}_{[k]}(\theta_{[k]}) = \mathbb{E}(\ell_{[k]}(\theta_{[k]}|X_1))$$

be the population MLE that $\hat{\theta}_{[k]}$ is estimating. Note that different models have a different sets of parameter, so the population MLE will be different.

Now we consider nested models:

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_K \tag{4.29}$$

in the sense that the number of parameters $d_1 < d_2 < \cdots < d_K$ are distinct in each model. Assume that there exists $k^*$ such that

$$X_1, \cdots, X_n \sim p_{\theta^*_{[k^*]}}. \tag{4.30}$$

Namely, $\mathcal{M}_{k^*}$ is the (minimal) correct model. Note that under the nested model assumption (equation (4.29)), $k^*$ model being correct implies that $\mathcal{M}_{k^*+1}, \cdots, \mathcal{M}_K$ are also correct model.

With the above notations, we can formally state the model selection consistency of BIC.

**Theorem 4.14** *Consider the model selection problem described in the above. Assume*

- **Nested model.** *The models are nested in the sense of equation (4.29) and $d_1 < d_2 < \cdots < d_K$.*
- **Correct model.** *The data is generated from equation (4.30), i.e., $\mathcal{M}_{k^*}$ is the minimal correct model.*

- **QMD for every model.** *Conditions (Q1-Q4) in Theorem 3.14 (asymptotic normality of M-estimator under QMD) holds for every model at its population MLE* $\theta^*_{[k]}$.

*Let* $\hat{k}$ *be the model selected by the BIC as in equation* (4.28)*. Then we have*

$$P(\hat{k} = k^*) \to 1.$$

**Proof:**

The idea of the proof is actually very simple. We consider two cases: models of a lower order $k < k^*$ and models of a higher order $k > k^*$.

**Outline of the proof.** The high-level idea is as follows.

- When $k < k^*$, we are using a wrong model, so we expect a constant gap $\Delta_k = \bar{\ell}_{[k^*]}(\theta^*_{[k^*]}) - \bar{\ell}_{[k]}(\theta^*_{[k]}) > 0$. So the BIC difference will be $BIC_{n,k} - BIC_{n,k^*} \approx \Delta_k \cdot n - (d_{k^*} - d_k) \log n$ diverges. Thus, asymptotically we will not choose $\mathcal{M}_k$.
- When $k > k^*$, all models are correct. So the difference $\bar{\ell}_{[k^*]}(\theta^*_{[k^*]}) - \bar{\ell}_{[k]}(\theta^*_{[k]}) = 0$ since they are all under correct model and the sample likelihood values $\ell_{n,[k^*]}(\theta^*_{[k^*]} \approx \ell_{n,[k^*]}(\theta^*_{[k^*]}$. Thus, $BIC_{n,k} - BIC_{n,k^*} \approx (d_k - d_{k^*}) \log n$ diverges. So we will choose $k^*$ rather than any $k$.

**Case** $k < k^*$**.** First, we want to note that $\ell_{n,[k]}(\theta_{[k]}) = \sum_{i=1}^n \ell_{[k]}(\theta_{[k]}|X_i)$ is of the order of $O_P(n)$ since it does not involves the factor $\frac{1}{n}$.

Once we divide it by $n$, we have the following result under the QMD conditions

$$\sup_{\theta_{[k]}} \left| \frac{1}{n} \ell_{n,[k]}(\theta_{[k]}) - \bar{\ell}_{[k]}(\theta_{[k]}) \right| = o_P(1). \tag{4.31}$$

This is known as Glivenko-Cantelli theorem (uniform convergence in probability, also known as uniform law of large numbers), which you will formally learn it in STAT 582-583. Equation (4.31) implies that

$$\begin{aligned} |\ell_{n,k} - n \cdot \bar{\ell}_{[k]}(\theta^*_{[k]})| &\equiv |\ell_{n,[k]}(\hat{\theta}_{[k]}) - n \cdot \bar{\ell}_{[k]}(\theta^*_{[k]})| \\ &\le n \left| \frac{1}{n} \ell_{n,[k]}(\hat{\theta}_{[k]}) - \bar{\ell}_{[k]}(\hat{\theta}_{[k]}) \right| + n \left| \bar{\ell}_{[k]}(\hat{\theta}_{[k]}) - \bar{\ell}_{[k]}(\theta^*_{[k]}) \right| \\ &\overset{(4.31)}{\le} o_P(n) + O_P\left( n \|\hat{\theta}_{[k]} - \theta^*_{[k]}\| \right) \\ &= o_P(n). \end{aligned} \tag{4.32}$$

Since $k < k^*$ implies that the model is incorrect, there exists a constant gap $\Delta_k = \bar{\ell}_{[k^*]}(\theta^*_{[k^*]}) - \bar{\ell}_{[k]}(\theta^*_{[k]}) > 0$. Thus,

$$\begin{aligned} BIC_{n,k} - BIC_{n,k^*} &= (d_k - d_{k^*}) \log n - 2(\ell_{n,k} - \ell_{n,k^*}) \\ &\overset{(4.32)}{=} (d_k - d_{k^*}) \log n - 2n \cdot (\bar{\ell}_{[k]}(\theta^*_{[k]}) - \bar{\ell}_{[k^*]}(\theta^*_{[k^*]})) + o_P(n) \\ &\ge (d_k - d_{k^*}) \log n + 2n \cdot \Delta_k + o_P(n), \end{aligned}$$

which diverges in probability. Thus, $P(\hat{k} < k^*) \to 0$.

**Case $k > k^*$.** Since all models are correct in this case, we immediately have that at the population MLEs $\theta^*_{[k]}$ and $\theta^*_{[k^*]}$, the likelihood value is always the same

$$\ell_{[k]}(\theta^*_{[k]}|x) = \log p_{\theta^*_{[k]}}(x) = \log p_{\theta^*_{[k^*]}}(x) = \ell_{[k^*]}(\theta^*_{[k^*]}|x).$$

Therefore,

$$\ell_{n,[k]}(\theta^*_{[k]}) = \sum_{i=1}^n \ell_{[k]}(\theta^*_{[k]}|X_i) = \sum_{i=1}^n \ell_{[k^*]}(\theta^*_{[k^*]}|X_i) = \ell_{n,[k^*]}(\theta^*_{[k^*]}). \tag{4.33}$$

Now we consider the sample MLE $\hat{\theta}_{[k]}$. Since it solves the first-order condition $\nabla \ell_{n,[k]}(\hat{\theta}_{[k]}) = 0$, the QMD condition implies the existence of a second-order Talyor expansion:

$$\begin{aligned}
\ell_{n,[k]}(\theta^*_{[k]}) - \ell_{n,[k]}(\hat{\theta}_{[k]}) &= \left(\theta^*_{[k]} - \hat{\theta}_{[k]}\right)^T \nabla\nabla \ell_{n,[k]}(\hat{\theta}^*_{[k]}) \left(\theta^*_{[k]} - \hat{\theta}_{[k]}\right) + o_P(n\|\theta^*_{[k]} - \hat{\theta}_{[k]}\|^2) \\
&= \sqrt{n}\left(\theta^*_{[k]} - \hat{\theta}_{[k]}\right)^T \nabla\nabla \frac{1}{n}\ell_{n,[k]}(\hat{\theta}^*_{[k]}) \sqrt{n}\left(\theta^*_{[k]} - \hat{\theta}_{[k]}\right) + o_P(n\|\theta^*_{[k]} - \hat{\theta}_{[k]}\|^2) \\
&= u_n^T \Gamma_n u_n + o_P(1),
\end{aligned}$$

where we know that

$$u_n\sqrt{n}\left(\theta^*_{[k]} - \hat{\theta}_{[k]}\right) = O_P(1)$$

since it converges in distribution and

$$\Gamma_n = \nabla\nabla \frac{1}{n}\ell_{n,[k]}(\hat{\theta}^*_{[k]}) = \Gamma + o_P(1)$$

since it converges in probability to a fixed matrix.

Therefore, we conclude that

$$\ell_{n,[k]}(\theta^*_{[k]}) - \ell_{n,[k]}(\hat{\theta}_{[k]}) = O_P(1).$$

Now we consider the difference in the BIC

$$\begin{aligned}
BIC_{n,k} - BIC_{n,k^*} &= (d_k - d_{k^*})\log n - 2(\ell_{n,k} - \ell_{n,k^*}) \\
&= (d_k - d_{k^*})\log n - 2(\ell_{n,[k]}(\hat{\theta}_{[k]}) - \ell_{n,[k^*]}(\hat{\theta}_{[k^*]})) \\
&= (d_k - d_{k^*})\log n - 2(\ell_{n,[k]}(\theta^*_{[k]}) - \ell_{n,[k^*]}(\theta^*_{[k^*]})) + O_P(1) \\
&\overset{(4.33)}{=} (d_k - d_{k^*})\log n + O_P(1).
\end{aligned}$$

Thus, the constant gap $(d_k - d_{k^*})\log n$ eventually exceed $O_P(1)$, so $P(\hat{k} > k^*) \to 0$, which completes the proof. ■

**Remark 4.15** *In the proof of BIC model consistency, you see that we are not limited to the BIC penalty $d\log n$. We just need the penalty to be $dr_n$ with any $r_n \to \infty$ and $r_n = o(n)$. The proof of BIC also implies that if we use AIC, we only have*

$$P(\hat{k}_{AIC} \geq k^*) \to 1,$$

*the one-sided model selection consistency. If our goal is to select a model with a good prediction performance that we are fine with a little bit of overfitting, then AIC is also a fine criterion. This also aligns with the derivation of AIC–the motivation is to estimate the empirical risk, so it avoids the underfitting but does not pay much attention to overfitting.*