Stat213 Lecture Notes

James Bailie

June 16, 2021

Abstract

These are my lecture notes for Stat213, a second graduate course in statistical inference, lectured by Prof. Pragya Sur in Spring 2021 at Harvard. All errors are my own. Sections marked add-on were not in the lecture and were added by me at a later point. Some diagrams are courtesy of Pragya Sur.

Contents

1 Lecture 26/1

1.1 General Framework

1. We are given observations $y_1, \ldots, y_n$ which are components of $\boldsymbol{y} \in \mathbb{R}^{dn}$.

2. Our goal is to learn about the underlying data generation process.

3. We assume that the observations are iid realisation of a random variable: $Y_1, \ldots, Y_n \overset{iid}{\sim} p^*$ with support $\mathscr{Y}$.

4. The general approach is this course can be summarised as follows:

   (a) Introduce a model – that is, a class $\mathcal{M} = \{p_\theta : \theta \in \mathcal{H}\}$ of probability distributions parametrised by $\theta \in \mathcal{H}$.

   (b) Assume that the model is well-specified – that is, that the true data generating process p^* is in the model class. We will also consider misspecified models, where $p^* \notin \mathcal{M}$.

   (c) In the well-specified setting, we wish to infer which $p_\theta \in \mathcal{M}$ equals p^*. In the misspecified setting, the goal is to find $p_\theta \in \mathcal{M}$ which is "closest" to p^* (in a sense that we will make precise later). In either setting, we call this θ the 'underlying θ'.

(d) To find the underlying θ, we design estimators $\hat{\theta}$ based on various techniques.

One technique for designing estimators is the method of moments. This technique was originally proposed when we had little computation resources and has now resurfaced as we have too much data to do hard computation. Another is likelihood based inference, which designs estimators based on the (log) likelihood function: given $Y_1, \dots, Y_n \overset{iid}{\sim} p_\theta$, the likelihood and log likelihood are (random) functions:

$$\begin{aligned} \mathcal{L}_n : \theta &\mapsto \prod_{i=1}^{n} p_\theta(Y_i), \\ l_n : \theta &\mapsto \sum_{i=1}^{n} \log p_\theta(Y_i), \end{aligned}$$

The standard likelihood based inference estimator is the maximum likelihood estimator (MLE), defined as either $\hat{\theta}_n^{\mathrm{ML}} = \operatorname{argmax}_{\theta \in \mathcal{H}} l_n(\theta)$ or as a solution to $\nabla_\theta l_n(\theta) = 0$. (Notation: θ may be a vector and ∇_θ is the gradient with respect to θ.)

Two other 'pseudo-likelihood' based techniques which generalise the MLE are M- and Z-estimators.

1.2 M- and Z-estimators

Definition 1.1. An *M-estimator* is obtained by maximising a (given) general criterion function

$$\theta \mapsto M_n(\theta) = \sum_{i=1}^{n} m_\theta(Y_i).$$

The M stands for maximise. M_n is only of the form $\sum_{i=1}^{n} m_\theta(Y_i)$ when the data are iid.

Example 1.2.

1. The MLE is an M-estimator where $m_\theta(Y_i) = \log p_\theta$.

2. Generalised (including non-linear) regression can be framed as an M-estimator problem. Given observations $(X_i, Z_i)_{i=1}^n$ such that $Z_i = f_\theta(X_i) + \varepsilon_i$ for some $\theta \in \mathcal{H}$ and iid ε_i. The least squares estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathcal{H}} \sum_{i=1}^{n} \left(Z_i - f_\theta(X_i)\right)^2,$$

is an M-estimator with $m_\theta(X_i, Z_i) = -\left(Z_i - f_\theta(X_i)\right)^2$.

The likelihood function can often be very complex or practically impossible to specify. An advantage of the M-estimator is that we do not need to consider or even specify a likelihood function.

Definition 1.3. A *Z-estimator* is defined as a solution to equations of the form

$$\sum_{i=1}^{n} \phi_\theta(Y_i) = 0 \tag{1}$$

for given functions ϕ_θ. The equation (1) is called the estimator equation(s).

The Z stands for zero. Some estimators are more easily expressed as M-estimators than Z-estimators and vice versa.

Example 1.4.

1. The MLE is a Z-estimator with $\phi_\theta = \nabla_\theta l_n(\theta)$.

2. Any M-estimator with m_θ differentiable is a Z-estimator with $\phi_\theta = \nabla_\theta m_\theta$.

3. The sample mean and sample median can be expressed as Z-estimator with $\phi_\theta(y) = y - \theta$ and $\phi_\theta(y) = \operatorname{sign}(y - \theta)$ respectively.

4. The Huber estimators (also known as the trimmed sample means) are Z-estimators with

$$\phi_\theta(y) = \begin{cases} k & \text{if } y - \theta \geq k, \\ -k & \text{if } y - \theta \leq -k, \\ y - \theta & \text{if } |y - \theta| < k. \end{cases}$$

The limiting function lim $k \to \infty\phi_\theta$ recovers the sample mean. Similarly $\lim_{k\to 0}\phi_\theta$ recovers the sample median. For $k \in (0, \infty)$, the result Z-estimator is a trimmed mean (which is robust to outliers for smaller k). The Huber estimators can also be expressed as M-estimators:

$$\hat{\theta} = \text{argmin} \sum_{i=1}^{n} m(Y_i - \theta),$$

where $m(y) = y^2 \mathbb{1}\{|y| \le k\} + (2k|y| - k^2)\,\mathbb{1}\{|x| > k\}$.

5. An example of an estimator problem that can be expressed as an M-estimator but not as a Z-estimator:

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Unif}(0, \theta).$$

(Most of the time if m_θ is not differentiable with respect to θ, then we have an M-estimator which cannot be written as a Z-estimator.)

1.3 Some model and estimator properties

The inference pipeline starts with a hypothesised model. The statistician then constructs a "good" estimator and does inference based on the estimators properties and the observed data.

What do we mean by a "good" estimator? It should satisfy at least two basic properties: I) It generates estimates which are close to the underlying θ, at least when given access to infinite samples. II) We should be able to provide rigorous uncertainty quantification for the estimator.

Definition 1.5. A model $\mathcal{M} = \{p_\theta : \theta \in \mathcal{H}\}$ is *well-specified* if there exists $\theta^* \in \mathcal{H}$ such that the true data generating process p^* equals p_{θ^*}. Note that θ^* does not have to be unique (in which case the model is unidentifiable). The goal in this case is to infer about θ^*.

When no $p_\theta \in \mathcal{H}$ equals p^*, then the model $\mathcal{M}$ is *misspecified.*

Definition 1.6. In the well-specified setting with θ^* unique, the estimator $\hat{\theta}_n$ is *consistent* if

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}^*} \theta^*,$$

where the notation $\mathbb{P}^*$ means convergence in probability under the true p^*.

Consistency is the formal definition of property I. What about property II? How do we characterise the fluctuation properties of $\hat{\theta}_n$? If $\hat{\theta}_n$ is consistent, then one measure of the fluctuation properties is the rate of convergence of $\hat{\theta}_n \to \theta^*$. The rate of convergence r_n is defined such that

$$r_n(\hat{\theta}_n - \theta^*)$$

converges to some distribution $\mathcal{F}$ (in this class $\mathcal{F}$ will always be Gaussian). r_n must blow up to infinity so as to couteract $(\hat{\theta}_n - \theta^*) \xrightarrow{\mathbb{P}^*} 0$. Also, $r_n(\theta_n - \theta^*)$ doesn't blow up. So r_n is large enough to stop $r_n(\theta_n - \theta^*) \to 0$ and small enough to stop $|r_n(\theta_n - \theta^*)| \to \infty$. Hence r_n can be thought of as the rate of convergence.

$r_n = \sqrt{n}$ is consider to be a fast rate of convergence. A good estimator will have $r_n \geq \sqrt{n}$.

In this course, we will look at estimators with $r_n(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{V}^*)$ where $\boldsymbol{V}^*$ is positive semidefinite. In the univariate case, if v_n is a consistent estimator for v^*, then we have a natural CI estimator

$$\hat{c}_n = \hat{\theta}_n \pm z_{\alpha/2}\sqrt{\frac{v_n}{r_n}},$$

with $\mathbb{P}^*(\hat{c}_n \ni \theta^*) \xrightarrow{n\to\infty} 1 - \alpha$.

1.3.1 The misspecified model setting

In the misspecified model setting there is no notion of a true parameter θ^*. The previous definition of consistency doesn't make sense in this case.

We need a new notion of consistency. We would still hope that even in the misspecified setting our estimator (such as the MLE) converges and its limit is meaningful. One reason for hope is that we can still define M- and Z-estimators in the misspecified setting.

For M-estimators, suppose that the general criterion function converges

$$M_n(\theta) \xrightarrow{\mathbb{P}^*} M^*(\theta), \tag{2}$$

as $n \to \infty$. (Note that $M_n(\theta)$ is a random variable since it is a function of $Y_1, \dots, Y_n$.)

Definition 1.7. Supposing equation (2) holds, then the corresponding M-estimator $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{\mathbb{P}^*} \theta^*$ where

$$\theta^* = \operatorname*{argmax}_{\theta \in \mathcal{H}} M^*(\theta).$$

Similarly for Z-estimators, if $\phi_{\theta,n} \xrightarrow{\mathbb{P}^*} \phi^*_\theta$, then the corresponding Z-estimator $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{\mathbb{P}^*} \theta^*$ where θ^* is the solution to $\phi^*_\theta(\boldsymbol{Y}) = 0$.

2 Lecture 28/1

2.1 Method of moments

Method of moments (MoM) defines estimators by matching sample and population moments. But in general, we do not need to restrict ourselves to moments. We can obtain estimators by matching any sample and population quantity; in this case, we have generalised method of moments estimators (although we will usually omit the term 'generalised').

Definition 2.1. Let $Y_1, \dots Y_n \overset{iid}{\sim} p_{\theta^*}$ with $\theta^* \in \mathcal{H}$. (So we are only defining the MoM estimators in the well specified setting.) A *(generalised) method of moment estimator* $\hat{\theta}$ is obtained by solving the system of equations

$$\frac{1}{n}\sum_{i=1}^{n} f_j(Y_i) = \mathbb{E}_\theta f_j(Y),$$

in terms of θ, where $f_1, \dots, f_k$ are some chosen functions and $Y \sim p_\theta$.

Typically if $\theta \in \mathbb{R}^d$ then $k = d$¿. The standard method of moments use the functions $f_j(y) = y^j$.

Remark 2.2. One advantage of the method of moments estimators is that we do not need to specify a criterion function. Another is that certain well known estimators can be expressed as MoM estimators. There is a class of models where the MLEs are MoM estimators. (We will see this later.) Moreover, MoM estimators have good properties in the limit of large samples, thanks to the delta method.

2.1.1 A central limit theorem for method of moments estimators

Theorem 2.3 (the delta method)**.** *Suppose $\hat{\theta}_n$ is a consistent estimator for $\theta \in \mathbb{R}^d$ with*

$$r_n(T_n - \theta) \xrightarrow{d} \mathcal{F}, \tag{3}$$

for some distribution $\mathcal{F}$. (Note that equation (3) implies that $\hat{\theta}_n$ is consistent.) Suppose $\phi : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at θ. Then

1. *$r_n(\phi(\hat{\theta}_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)\mathcal{F}$, where $\phi'(\theta) \in \mathbb{R}^{k \times d}$ is the Jacobian of ϕ evaluated at θ:*
$$[\phi'(\theta)]_{ij} = \frac{\partial \phi_i(\theta)}{\partial \theta_j}.$$

2. $$r_n(\phi(\hat{\theta}_n) - \phi(\theta)) - r_n\phi'(\theta)(T_n - \theta) \xrightarrow{P} 0.$$

The proof is left as an exercise. (It is a simple application of Taylor's theorem using Slutsky's lemma and the continuous mapping theorem.)

Example 2.4. If $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(\mu, \Sigma)$ and ϕ satisfies the theorem assumptions then

$$\sqrt{n}\left(\phi(T_n) - \phi(\theta)\right) \xrightarrow{d} \mathcal{N}\left(\phi'(\theta)\mu, \phi'(\theta)\Sigma\phi'(\theta)^\mathsf{T}\right).$$

So we automatically get asymptotic normality of $\phi(T_n)$ from asymptotic normality of T_n!

Definition 2.5 (some notation from empirical process theory)**.** Given a distribution (i.e. a probability measure) P on $\mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}^d$, define

$$Pf := \int f dP = \int_{\mathcal{X}} f(x) dP(x) = \mathbb{E}_p[f(X)].$$

Given observations $x_1, \dots, x_n$ from P, define

$$P_n f = \frac{1}{n} \sum_{i=1}^{n} f(x_i),$$

to be the empirical PDF of $f(X)$.

Recall that a MoM estimator $\hat{\theta}_n$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} f_j(Y_i) = \mathbb{E}_{\hat{\theta}_n}[f_j(Y)], \tag{4}$$

for $j = 1, \dots, k$. Define $\underset{\sim}{f} = (f_1, \dots, f_k)$ and $e(\theta) := \mathbb{P}_\theta \underset{\sim}{f}$ where $\mathbb{P}_\theta$ is the probability measure under p_θ. With this notation, equation (4) becomes

$$\mathbb{P}_n \underset{\sim}{f} = e(\hat{\theta}_n).$$

When does a solution to equation (4) exist? A necessary condition is that $\mathbb{P}_n \underset{\sim}{f}$ is in the range of e. Further, if we suppose that e is one-to-one, then $\hat{\theta}_n = e^{-1}(\mathbb{P}_n \underset{\sim}{f})$ is the unique MoM estimator.

From the central limit theorem, we know that

$$\sqrt{n}\left(\mathbb{P}_n \underset{\sim}{f} - \mathbb{P}_\theta \underset{\sim}{f}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \operatorname{Cov}(\underset{\sim}{f}(Y)) = \mathbb{E}_\theta\left[\underset{\sim}{f}(Y)\underset{\sim}{f}(Y)^{\mathsf{T}}\right] - \left[\mathbb{E}_\theta \underset{\sim}{f}(Y)\right]\left[\mathbb{E}_\theta \underset{\sim}{f}(Y)\right]^{\mathsf{T}} = \mathbb{P}_\theta f f^{\mathsf{T}} - \mathbb{P}_\theta f (\mathbb{P}_\theta f)^{\mathsf{T}}.$$

The delta method gives an asymptotic distribution of the (unique) MoM estimator $\hat{\theta}_n$:

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(e^{-1}\left(\mathbb{P}_n \underset{\sim}{f}\right) - e^{-1}\left(\mathbb{P}_\theta \underset{\sim}{f}\right)\right) \\
&\xrightarrow{d} \left[e^{-1}\left(\mathbb{P}_\theta f\right)\right]' \mathcal{N}(0, \Sigma) \\
&= \mathcal{N}\left(0, \left[e^{-1}\left(\mathbb{P}_\theta f\right)\right]' \Sigma \left(\left[e^{-1}\left(\mathbb{P}_\theta f\right)\right]'\right)^{\mathsf{T}}\right),
\end{aligned}$$

where $\left[e^{-1}\left(\mathbb{P}_\theta f\right)\right]'$ is the Jacobian of e^{-1} evaluated at $\mathbb{P}_\theta f$ (not the derivative of θ). That is, $\left[e^{-1}\left(\mathbb{P}_\theta f\right)\right]' = \left.(e^{-1})'\right|_{\mathbb{P}_\theta f}$.

Lemma 2.6 (the inverse function theorem)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable in a neighbourhood of $\theta \in \mathbb{R}^d$. Suppose that $F'|_{\theta=\theta_0} \in \mathbb{R}^{d\times d}$ is invertible. Then F^{-1} is well defined and continuously differentiable in some neighbourhood $N_{F(\theta_0)}$ of $F(\theta_0)$. Moreover,*

$$\left[F^{-1}\right]'\Big|_t = \left[F'\left(F^{-1}(t)\right)\right]^{-1}, \tag{5}$$

for $t \in N_{F(\theta_0)}$.

The LHS of equation (5) is the Jacobian of F^{-1} evaluated at t. The RHS is the inverse of {the Jacobian of F evaluated at $F^{-1}(t)$}.

If e satisfies Lemma 2.6 then

$$\left[e^{-1}\left(\mathbb{P}_\theta f\right)\right]' = \left[e'(\theta)\right]^{-1},$$

since $e(\theta) = \mathbb{P}_\theta \underset{\sim}{f}$ so that $e^{-1}(\mathbb{P}_\theta \underset{\sim}{f}) = \theta$. The RHS is the inverse of {the Jacobian of e evaluated at θ}.

The following theorem recaps the above discussion.

Theorem 2.7. *Let $e(\theta) = \mathbb{P}_\theta \underset{\sim}{f}$ be one-to-one on an open set $\Theta \subset \mathbb{R}^d$ and continuously differentiable at $\theta^* \in \Theta$. Assume that $e'(\theta^*)$ is non-singular; the L^2 norm $\mathbb{P}_{\theta^*}\left\|\underset{\sim}{f}\right\|^2$ is finite (so that we can apply the CLT); and $Y_i \overset{iid}{\sim} P_{\theta^*}$ (so we are in the well specified setting). Then the MoM estimator $\hat{\theta}_n = e^{-1}\left(\mathbb{P}_n \underset{\sim}{f}\right)$ exists with probability going to 1*[*] *(as $n \to \infty$) and*

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) &\xrightarrow{d} \mathcal{N}\left(0, \left[e'(\theta^*)\right]^{-1} \mathrm{Cov}_{\theta^*}(\underset{\sim}{f}) \left(\left[e'(\theta^*)\right]^{-1}\right)^{\mathsf{T}}\right) \\
&= \mathcal{N}\left(0, \left[e'(\theta^*)\right]^{-1} \left[\mathbb{P}_{\theta^*}\underset{\sim}{f}\underset{\sim}{f}^{\mathsf{T}} - e(\theta^*)e(\theta^*)^{\mathsf{T}}\right] \left(\left[e'(\theta^*)\right]^{-1}\right)^{\mathsf{T}}\right) \\
&= \mathcal{N}\left(0, \left[e'(\theta^*)\right]^{-1} \mathbb{P}_{\theta^*}\underset{\sim}{f}\underset{\sim}{f}^{\mathsf{T}} \left(\left[e'(\theta^*)\right]^{-1}\right)^{\mathsf{T}}\right),
\end{aligned}$$

where the convergence is under $\mathbb{P}_{\theta^*}$[†]*.*

So we have obtained a CLT for MoM estimators just from using the delta method. For other estimators (such as M- and Z-estimators) we will need more theory.

[*] The MoM estimator exists as soon as $\mathbb{P}_n \underset{\sim}{f} \in e(\Theta)$ which happens with probability tending to one by the WLLN.

[†] add-on Apparently we can cancel the term $e(\theta^*)e(\theta^*)^{\mathsf{T}}$ but I don't understand why.

2.1.2 MoM Estimators in the high dimensional setting

MoM estimators are useful in modern research in the high dimensional supervised setting. See the survey paper that Pragya references. In this setting, we have observations of the outcome Y and the covariates $\boldsymbol{X}$. The number of samples n goes to infinity and the number of features p also goes to infinite at a rate comparable to or faster than n (that is, $p \neq o(n)$). In this case, we typically must have $\dim \underset{\sim}{f} \to \infty$.

2.2 The exponential families

Definition 2.8. A class of distributions $\{p_\theta : \theta \in \mathcal{H}\}$ with $\mathcal{H} \subset \mathbb{R}^d$ is an *exponential family* if there exists a sufficient statistic $T : \mathcal{X} \to \mathbb{R}^d$ such that the density p_θ factors as

$$p_\theta(x) = h(x) \exp\left[\theta^\mathsf{T} T(x) - A(\theta)\right] d\mu(x),$$

with regard to the base measure μ and where h and μ do not depend on θ and

$$A(\theta) = \log \int h(x) \exp\left[\theta^\mathsf{T} T(x)\right] d\mu(x),$$

is the integrating constant. A is called the *cumulant function* and θ is called the *natural parameter*. The *natural parameter space* Θ is the set of parameters θ for which p_θ exists (i.e for which $A(\theta) < \infty$). If Θ is non-empty and open then we say that the family is *regular*.

2.2.1 Some fundamental properties

The moments of the sufficient statistic can be linked to the derivatives of the cumulant.

Lemma 2.9. *$A(\theta)$ is convex and infinitely differentiable with*

$$\frac{\partial^k e^{A(\theta)}}{\partial\theta_1^{\alpha_1} \dots \partial\theta_d^{\alpha_d}} = \int h(x) T_1(x)^{\alpha_1} \dots T_d(x)^{\alpha_d}, \tag{6}$$

where $\sum_{j=1}^d \alpha_j = k$ and $\alpha_j \in \mathbb{N}$. (To prove equation (6) we need only prove that we can exchange derivatives and integrals.)

(add-on) I think this lemma shows that A is the cumulant generating function.

We omit the proof of this lemma. We have

$$\nabla_\theta e^{A(\theta)} = e^{A(\theta)} \nabla_\theta A(\theta) = \left(\int_{\mathcal{X}} h(x) T_j(x) \exp\left(\theta^{\mathsf{T}} T(x)\right) d\mu(x) \right)_{j=1,\ldots,d}$$

so that

$$\nabla_\theta A(\theta) = \mathbb{E}_\theta[T(X)]. \tag{7}$$

We can also show that the Hessian

$$\nabla_\theta^2 A(\theta) = \operatorname{Cov}(T). \tag{8}$$

2.2.2 MLE in exponential families

The MLE maximises $l_n(\theta) - h(x) = \theta^{\mathsf{T}} \mathbb{P}_n T - A(\theta)$. (Recall the notation $\mathbb{P}_n f$ from Definition 2.5.) Differentiating with regard to θ, the MLE is a solution to the equation

$$\mathbb{P}_n T - \nabla_\theta A(\theta) = 0.$$

Yet $\nabla_\theta A(\theta) = \mathbb{E}_\theta[T(X)]$ (equation 7). Hence the MLE satisfies $\mathbb{P}_n T = \mathbb{E}_\theta[T(X)]$. This equation should look familiar from Theorem 2.7 – the MLE is a method of moments estimator with $f = T$. If $e(\theta) = \mathbb{E}_\theta[T(x)]$ satisfies the conditions of Theorem 2.7, then the MLE is given by

$$\hat{\theta}_n = e^{-1}\left(\mathbb{P}_n T\right),$$

and

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_n - \theta^*\right) &\xrightarrow{d} \mathcal{N}\left(0, \left[e'(\theta^*)\right]^{-1} \operatorname{Cov}_{\theta^*}\left[T(X)\right] \left(\left[e'(\theta^*)\right]^{-1}\right)^{\mathsf{T}}\right) \\ &= \mathcal{N}\left(0, \left(\operatorname{Cov}_{\theta^*}\left[T(X)\right]\right)^{-1}\right) \end{aligned}$$

where the convergence is under $\mathbb{P}_{\theta^*}$ and the second line follows from the fact that $e(\theta) = \nabla_\theta A(\theta)$ so that $e'(\theta) = \nabla_\theta^2 A(\theta)|_{\theta=\theta^*} = \operatorname{Cov}_{\theta^*}\left[T(X)\right]$ by equation 8.

Definition 2.10. Note that for the assumptions of Theorem 2.7 to hold, we need that $\operatorname{Cov}_{\theta^*}\left[T(X)\right]$ is non-singular. If $\operatorname{Cov}_{\theta^*}\left[T(X)\right]$ is non-singular then we call the exponential family *full*.

Can we represent the asymptotic covariance in another form using the likelihood function? Yes! We know that $\nabla_\theta l_n(\theta) = T(x) - \nabla_\theta A(\theta)$ and so $\nabla_\theta^2 l_n(\theta) = -\nabla_\theta^2 A(\theta) = -\mathrm{Cov}_\theta[T(X)]$. Since $\nabla_\theta^2 l_n(\theta)$ doesn't depend on X, we have

$$\nabla_\theta^2 l_n(\theta) = \mathbb{E}_\theta\left[\nabla_\theta^2 l_n(\theta)\right] = -I(\theta),$$

where $I(\theta)$ is the Fisher information. So in the case of full exponential families (with e satisfying the assumptions of Theorem 2.7), we get the classical MLE asymptotic theorem

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

for free, after establishing the MoM asymptotic Theorem 2.7.

Example 2.11 (linear regression). Let $(x_i, y_i)_{i=1,\dots,n}$ be paired observations with $y_i = x_i^\top \theta^* + \epsilon_i$ and $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. The log likelihood is $-1/2\|y - X\theta\|^2$. We can write the density as an expoential family and we can show that the MLE is given by

$$\hat{\theta}_n^{\mathrm{ML}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Using our asymptotic result, we get that

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_{\theta^*}\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]\right).$$

Two examples of exponential families where the assumption of this theory aren't satisfied:

1. When Cov_{θ^*} is singular;

2. In curved exponential families (for example $\{\mathcal{N}(\theta, \theta^2) : \theta \in \mathbb{R}\}$).

3 Lecture 2/2

3.1 Consistency of the MLE

We will work in the following set-up. Let $Y_1, \dots, Y_n \overset{iid}{\sim} p^*$. (Note that there is no parameter, so we are not necessarily in the well-specified setting.) Let $\mathcal{M} = \{p_\theta : \theta \in$

$\mathcal{H}\}$ be the hypothesised model class. The MLE is defined as

$$\hat{\theta}^{\mathrm{ML}} = \underset{\theta \in \mathcal{H}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_\theta(y_i).$$

To study the consistency of the MLE, we start with the question, what are the basic properties of the above criterion function? Fix $\theta \in \mathcal{H}$. By the SLLN (and the iid assumption), assuming that $\mathbb{E}_*|\log p_\theta(Y)| < \infty$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(Y_i) \xrightarrow[n \to \infty]{\mathbb{P}_*} \mathbb{E}_* \left[\log p_\theta(Y)\right],$$

where $\mathbb{P}_*$ is convergence in probability under the true data generating process p^* and $\mathbb{E}_*$ is expectation with respect to this p^*.

Define $l^* : \theta \mapsto \mathbb{E}_* \left[\log_\theta(Y)\right]$. We hope that the MLE, if it converges, converges to a maximiser of l^*. We will show later that this is true (under some conditions)!

Define $\theta^* = \operatorname{argmax}_{\theta \in \mathcal{H}} l^*(\theta)$. Note that previously we used the notation θ^* to denote the true data generating parameter; this is no longer necessarily the case (although we will see later that it is under mild assumptions in the well-specificed case). Instead θ^* is simply the maximiser of the theoretical log likelihood $\mathbb{E}_* \left[\log_\theta(Y)\right]$. How can we interpret θ^*?

3.1.1 θ^* in the well-specified setting

Let $\tilde{\theta} \in \mathcal{H}$ be such that $p_{\tilde{\theta}} = p_*$ is equal to the true data generating process. (So $\tilde{\theta}$ is the true data generating parameter – this notation replaces the previously used θ^*. We have also replaced the notation p^* with p_*.) We would hope that $\theta^* = \tilde{\theta}$. This is true under mild conditions, by the following Proposition.

Proposition 3.1 (Lemma 5.35 of [vdV]). *Let $\{p_\theta : \theta \in \mathcal{H}\}$ be a well specified model and $\tilde{\theta}$ be the data generating value. Assume that $p_\theta(y) > 0$ for all $\theta \in \mathcal{H}$ and all $y \in \mathcal{Y}$ and that the model is identifiable. Then the function*

$$\theta \mapsto l^*(\theta) = \int_{\mathcal{Y}} \log p_\theta(y) p_{\tilde{\theta}}(dy),$$

is maximised uniquely at $\tilde{\theta}$, the true data generating parameter.

Proof. Maximising l^* is equivalent to maximising

$$C(\theta) = \int_{\mathcal{Y}} \log\left[\frac{p_\theta(y)}{p_{\tilde{\theta}}(y)}\right] p_{\tilde{\theta}}(dy),$$

(since $C(\theta)$ is equal to $l^*(\theta)$ up to an additive constant $\int_{\mathcal{Y}} \log[p_{\tilde{\theta}}(y)]p_{\tilde{\theta}}(dy)$.)

Observe that $C(\tilde{\theta}) = 0$. Hence it suffices to show that $C(\theta) < 0$ for all $\theta \neq \tilde{\theta}$. We can use Jensen's inequality or (as we will do) use the fact that $\log x \leq 2(\sqrt{x} - 1)$ for all $x > 0$: For $\theta \neq \tilde{\theta}$,

$$\begin{aligned} C(\theta) &\leq 2\int_{\mathcal{Y}} \left(\sqrt{\frac{p_\theta(y)}{p_{\tilde{\theta}}(y)}} - 1\right) p_{\tilde{\theta}}(dy) \\ &= \int_{\mathcal{Y}} \sqrt{p_\theta(y)p_{\tilde{\theta}}(y)} - p_{\tilde{\theta}}(y) - p_\theta(y)dy \\ &= -\int_{\mathcal{Y}} \left(\sqrt{p_\theta(y)} - \sqrt{p_{\tilde{\theta}}(y)}\right)^2 dy \\ &< 0, \end{aligned}$$

where the second line follows from the fact that $\int_{\mathcal{Y}} p_{\tilde{\theta}}(y)dy = \int_{\mathcal{Y}} p_\theta(y)dy = 1$ and the last line follows by identifiability. This provees that the maximiser θ^* exists, is unique and equals $\tilde{\theta}$. □

3.1.2 θ^* in the misspecified setting

In this setting we replace C with the analogue

$$D : \theta \mapsto \int_{\mathcal{Y}} \log\left[\frac{p_\theta(y)}{p_*(y)}\right] p_*(dy),$$

where p_* is the true data generating distribution. Minimising D is equivalent to minimising the KL-divergence between p_* and p_θ:

$$\theta^* = \operatorname*{argmax}_{\theta \in \mathcal{H}} l^*(\theta) = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathrm{KL}(p_*||p_\theta),$$

where $\mathrm{KL}(p||q) = \int_{\mathcal{Y}} \log\left[\frac{p(y)}{q(y)}\right] p(dy)$. Hence p_{θ^*} is the KL projection of p_* onto $\mathcal{M}$.

In the well-specified case, we have $D(\theta^*) = 0$. In the misspecified case, we have

$$D(\theta^*) = \mathbb{E}_* \left(\log \left[\frac{p_\theta(y)}{p_*(y)} \right] \right) < \log \mathbb{E}_* \left(\frac{p_\theta(y)}{p_*(y)} \right) = 0.$$

Hence, $D(\theta^*)$ can be interpreted as a measure of how misspecified the model is.

3.1.3 Sufficient conditions for convergence to θ^*

Recap: Recall that $\theta^* = \operatorname{argmax}_{\theta \in \mathcal{H}} \mathbb{E}_* [\log p_\theta(Y)]$ is the maximiser of the theoretical log-likelihood; it is not necessarily the true data generating parameter. However, under mild conditions, we showed (Proposition 3.1) that θ^* is the true data generating parameter in the well-specified setting; and in the misspecified setting p_{θ^*} is the KL-projection of the true data generating parameter p_* onto the model space $\mathcal{M}$.

Suppose that $\frac{1}{n} l_n(\theta)$ converges pointwise (in θ) to $l^*(\theta)$ in $\mathbb{P}^*$-probability. (Note that we can view each $l_n(\theta)$ as a random variable, since it is a function of $Y_1, \dots, Y_n$.) Is this sufficient to guarantee $\hat{\theta}^{\mathrm{ML}} \to \theta^*$? No – here is a counterexample: Suppose $\mathcal{H}$ is infinite. Further suppose that $\frac{1}{n} l_n(\theta)$ has a spike at $\theta \neq \theta^*$ and the spike moves as n increases, so that for each θ, there is an n large enough so that there is no spike at $\frac{1}{n} l_n(\theta)$. Then we have pointwise convergence, but $\operatorname{argmax}_{\theta \in \mathcal{H}} \frac{1}{n} l_n(\theta)$ will follow the spike as $n \to \infty$ and not converge to θ_*. More concretely, let $\mathcal{H} = \mathbb{R}$ and suppose the spike moves off to infinity as $n \to \infty$.

This example shows that we need uniform convergence to guarantee consistency of the MLE.

Theorem 3.2 (consistency result #1)**.** *Let l_n be random functions on $\mathcal{H}$ (i.e. $l_n(\theta)$ is a random variable) and let l^* be a fixed function on $\mathcal{H}$. Let $\theta^* \in \mathcal{H}$ be the mode of l^*. Suppose the assumptions:*

1. *convergence of $l_n(\theta)$ to $l(\theta)$ which is uniform in θ – that is:*

$$\sup_{\theta \in \mathcal{H}} \left| \frac{1}{n} l_n(\theta) - l^*(\theta) \right| \xrightarrow[n \to \infty]{\mathbb{P}_*} 0;$$

2. *the mode of l^* is well separated: for all $\epsilon > 0$,*

$$\sup_{\substack{\theta \in \mathcal{H} \\ d(\theta, \theta^*) \geq \epsilon}} l^*(\theta) < l^*(\theta^*).$$

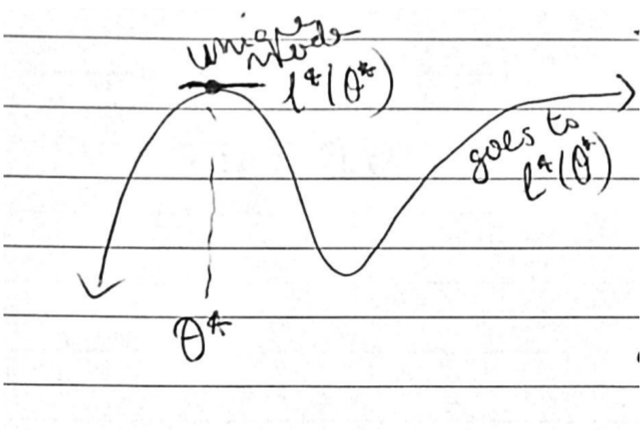


Figure 1: A situation prohibited by the well-separated mode assumption.

(Note that the supremum of an empty set is negative infinity.) Then any sequence $\hat{\theta}_n$ maximising l_n converges to θ^* in $\mathbb{P}_*$-probability. (Here $\mathbb{P}_*$-probability is the true generating process of the random variables $l_n(\theta)$.)

Note: throughout this course we will assume that all sets are at least metric spaces. This means that we can equip a metric d to $\mathcal{H}$. Assumption 2. implies that θ^* is a global mode of l^* but it is a stronger statement: it prohibits a sequence $\{l^*(\theta_n)\}$ from reaching $l^*(\theta^*)$ except when $\lim_{n\to\infty}\theta_n = \theta^*$ (for example, when $\mathcal{H} = \mathbb{R}$, $l^*(\theta)$ cannot increase to $l^*(\theta_*)$ as $\theta \to \infty$). See the section materials for an example of this. add-on This condition cannot be replaced with $\sup_{\substack{\theta\in\mathcal{H}\\ \theta\neq\theta}} < l^*(\theta^*)$ – such a change would force l^* to be discontinuous at θ^*.

Proof.

1. $\frac{1}{n}l_n(\theta^*) = l^*(\theta^*) + o_{\mathbb{P}_*}(1)$ by the convergence of l_n. (Here $o_{\mathbb{P}_*}(1)$ is a random variable that converges to zero in $\mathbb{P}^*$-probability.)
2. $\hat{\theta}_n$ maximises $\frac{1}{n}l_n(\theta)$ so that $\frac{1}{n}l_n(\hat{\theta}_n) \geq \frac{1}{n}l_n(\theta^*)$.

We get

$$l^*(\theta^*) - l^*(\hat{\theta}_n) = \frac{1}{n}l_n(\theta^*) + o_{\mathbb{P}_*}(1) - l^*(\hat{\theta}_n)$$

$$
\begin{aligned}
&\leq \frac{1}{n} l_n(\hat{\theta}_n) - l^*(\hat{\theta}_n) + o_{\mathbb{P}_*}(1) \\
&\leq \sup_{\theta \in \mathcal{H}} \left| \frac{1}{n} l_n(\theta) - l^*(\theta) \right| + o_{\mathbb{P}_*}(1) \\
&\xrightarrow[n\to\infty]{\mathbb{P}_*} 0,
\end{aligned}
$$

where the first line follows from 1.; the second from 2. and the third by uniform convergence (note that we need uniform convergence since we are bounding l_n at a random point $\hat{\theta}_n$). But we also know that

$$l^*(\theta^*) - l^*(\hat{\theta}_n) \geq 0,$$

$\mathbb{P}_*$-a.s. So $l^*(\hat{\theta}_n) \xrightarrow[n\to\infty]{\mathbb{P}_*} l^*(\theta^*)$.

Fix $\epsilon > 0$ and some $\eta = \eta(\epsilon) > 0$. As events,

$$
\begin{aligned}
\{d(\hat{\theta}_n, \theta^*) \geq \epsilon\} &\subseteq \{l^*(\hat{\theta}_n) \leq l^*(\theta^*) - \eta\} \\
&\subseteq \left\{ \left| l^*(\hat{\theta}_n) - l^*(\theta^*) \right| \geq \eta \right\}
\end{aligned}
$$

The probability of event $\left\{ \left| l^*(\hat{\theta}_n) - l^*(\theta^*) \right| \geq \eta \right\}$ goes to 0 by the previous reasoning. Hence $\mathbb{P}_* \left(d(\hat{\theta}_n, \theta^*) \geq \epsilon \right) \to 0$. □

Remark 3.3. We could modify the definition of $\hat{\theta}_n$ to be approximate maximisers: that is, we need only require

$$l_n(\hat{\theta}_n) \geq l_n(\theta) + o_{\mathbb{P}_*}(1),$$

for all $\theta \in \mathcal{H}$.

Definition 3.4.

1. The definition of *small-o in probability*: $X_n = o_{\mathbb{P}}(R_n)$ if there exists a random variable Z_n such that $X_n = Z_n R_n$ and $Z_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$.

2. The definition of *big-O in probability*: $X_n = O_{\mathbb{P}}(R_n)$ if there exists a random variable Z_n such that $X_n = Z_n R_n$ and Z_n is bounded in $\mathbb{P}$-probability – that is, for all $\epsilon > 0$, there is a constant M such that

$$\sup_{n \in \mathbb{N}} \mathbb{P}(\|Z_n\| > M) < \epsilon.$$

4 Lecture 4/2

4.1 Uniform convergence

Recall the definition of uniform convergence (assumption 1 in Theorem 3.2):

Definition 4.1. A sequence of random functions l_n with domain $\mathcal{H}$ *converges uniformly* to l^* if

$$\sup_{\theta \in \mathcal{H}} |l_n(\theta) - l^*(\theta)| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0.$$

Definition 4.2. Suppose $\mathcal{F}$ is a class of functions and $\{X_n\}_{n\in\mathbb{N}}$ are iid random variables with density p^*. The *uniform law of large numbers (ULLN)* holds for $\mathcal{F}$ and $\{X_n\}$ if

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_*[f(X)] \right| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0.$$

Example 4.3 (Glivenko-Cantelli theorem). Let $\{X_n\}_{n\in\mathbb{N}}$ be iid from a distribution with CDF F^*. Let

$$\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}\}.$$

Then the ULLN holds for $\mathcal{F}$ and $\{X_n\}$:

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i) - F^*(t) \right| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0.$$

That is, the empirical CDF converges uniformly to the true CDF in P_*-probability. (Proof: in section 2 we show a.s. convergence.)

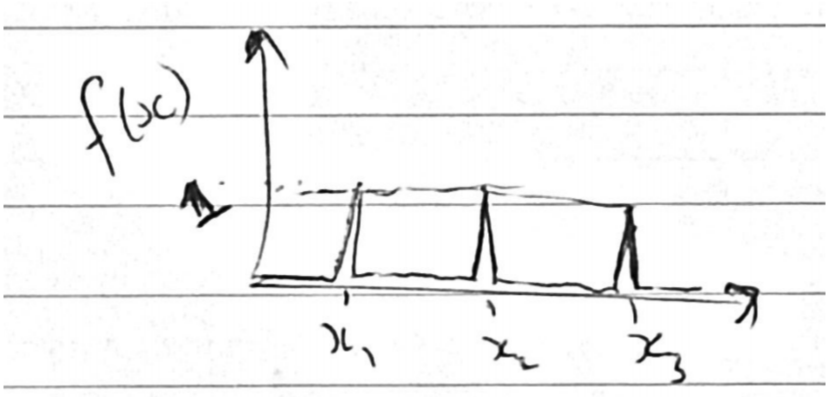
Consider continuous $\{X_n\} \overset{iid}{\sim} p^*$. Can you think of a property of $|mathcalF$ such that the ULLN will not be satisfied? One possible answer: whenever the space of functions $\mathcal{F}$ is too large. For example $\mathcal{F} = \{f | f : \mathbb{R} \to [0,1] \text{ continuous}\}$. Each f is bounded so the LLN holds for all $f \in \mathcal{F}$. But we will show that the ULLN doesn't hold for $\mathcal{F}$.

Take $\delta > 0, n \in \mathbb{N}$ and any $x_1, \ldots, x_n \in \mathbb{R}$. Then there exists $f \in \mathcal{F}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} f(x_i) \geq 1 - \delta$$

$$\mathbb{E}_\theta[f(X)] \leq \delta$$

The idea for constructing f is to set f to be zero, except it has peaks of 1 at $x_1, \ldots, x_n$, and then make it continuous.



We can construct such an f for every realisation of $X_1, \ldots, X_n$. Thus,

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_*[f(X)] \right| \geq 1 - 2\delta,$$

with $\mathbb{P}_*$-probability 1.

We will discuss ULLN more in the section. See also the paper (1979) *Empirical Processes: A survey of results for iid random variables.*

What are some sufficient conditions under which ULLN holds?

Proposition 4.4. *Let $\{q_\theta : \theta \in \mathcal{H}\}$ be a family of functions such that $q_\theta : \mathcal{Y} \subset \mathbb{R}^p \to \mathbb{R}$ is measurable and $Y_1, Y_2, \ldots \overset{iid}{\sim} p^*$ with support $\mathcal{Y}$. Assume*

1. *$\mathcal{H}$ is compact. (Also, throughout this course we assume that everything is at least a metric space.)*
2. *For all $y \in \mathcal{Y}$, the function $\theta \mapsto q_\theta(y)$ is continuous.*
3. *A domination condition (which captures the idea that the function class is not too large):*

$$\int \sup_{\theta \in \mathcal{H}} |q_\theta(y)| p^*(dy) < \infty.$$

Then

$$\sup_{\theta \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} q_\theta(Y_i) - \mathbb{E}_\theta \left[q_\theta(Y) \right] \right| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0.$$

(We don't actually need $\mathcal{Y} \subset \mathbb{R}^p$.)

Proof. See Question 8, Assignment 1. □

add-on The rest of this section is taken from Section 2. The proof of Proposition uses two lemmas (for their proofs see Section 2):

Lemma 4.5. *Let $\mathcal{C} = \{h_1, \ldots, h_K\}$ be a finite family of measurable functions from some space T to $\mathbb{R}$. Given $X_1, \ldots, X_n$ iid from some distribution such that $\mathbb{E}|h_i(X)| < \infty$ for $i = 1, \ldots, K$, we have a uniform strong law of large numbers (USLLN):*

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}[h(X_1)] \right| \xrightarrow[n\to\infty]{a.s.} 0.$$

Lemma 4.6. *Let $\mathcal{F}$ be a class of functions from some space T to $\mathbb{R}$. Assume that for every $\epsilon > 0$, there exists a finite number N_ϵ of brackets $[l_j, u_j]$ such that*

1. $\mathbb{E}_\star|l_j(X)| < \infty$ *and* $\mathbb{E}_\star|u_j(X)| < \infty$ *for all* $j = 1, \ldots, N_\epsilon$*;*
2. $\mathbb{E}_\star|u_j(X) - l_j(X)| < \epsilon$ *for all* $j = 1, \ldots, N_\epsilon$*;*
3. *for all $h \in \mathcal{F}$, there exists some j such that $h \in [l_j, u_j]$ – that is, $l_j(x) \leq h(x) \leq u_j(x)$ for all $x \in T$.*

Then a USLLN holds:

$$\sup_{h \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}[h(X)] \right| \xrightarrow[n\to\infty]{a.s.} 0.$$

4.2 Consistency of MLE (revisited)

Theorem 4.7 (consistency result #2)**.** *Let $\{p_\theta : \theta \in \mathcal{H}\}$ be a family of distributions with common support $\mathcal{Y}$. Suppose we are given observations $Y_1, \ldots, Y_n \overset{iid}{\sim} p^*$. Assume that*

1. $\mathcal{H}$ is compact.

2. (a) For all $\theta \in \mathcal{H}$ and $y \in \mathcal{Y}$, $p_\theta(y) > 0$.

(b) The function $y \mapsto p_\theta(y)$ is measurable for all $\theta \in \mathcal{H}$.

(c) The function $\theta \mapsto p_\theta(y)$ is continuous for all $y \in \mathcal{Y}$.

3. A domination condition:

$$\int \sup_{\theta \in \mathcal{H}} |\log p_\theta(y)| p^*(dy) < \infty.$$

4. The function

$$\theta \mapsto l^*(\theta) = E_\theta\left(\log[p_\theta(y)]\right) = \int \log[p_\theta(y)] p^*(dy)$$

is uniquely maximised at θ^*.

Then the MLE is consistent: $\hat{\theta}^{ML} \xrightarrow[n\to\infty]{\mathbb{P}_*} \theta^*$.

Proof. We get uniform convergence by Proposition 4.4. It suffices to show that: i) l^* is continuous; and ii) the mode θ^* of l^* is well separated. Then we can apply Theorem 3.2.

To prove i), we will show that if $\theta_n \to \theta_0$ then $l^*(\theta_n) \to l^*(\theta_0)$ – that is,

$$\mathbb{E}_*\left[\log p_{\theta_n}(Y)\right] \to \mathbb{E}_*\left[\log p_{\theta_0}(Y)\right]. \tag{9}$$

By 2c, we have that $\log p_{\theta_n}(Y) \to \log p_{\theta_0}(Y)$. Condition 3. gives us that the uniform dominator $g(Y) := \sup_{\theta\in\mathcal{H}} |\log p_\theta(Y)|$ is in L^1 so we can apply the DCT to get equation 9 as desired.

Now we will prove ii). We want to show that for all $\epsilon > 0$, there exists a constant $c > 0$ such that

$$\sup_{\substack{\theta\in\mathcal{H} \\ d(\theta,\theta^*)\geq\epsilon}} l^*(\theta) < l^*(\theta^*) - c.$$

We proceed by contradiction. Suppose there exists $\epsilon > 0$, a positive sequence $\{c_k\}$ converging to zero, and a sequence $\{\theta_k\} \subset \{\theta \in \mathcal{H} : d(\theta, \theta^*) \geq \epsilon\}$ such that

$$l^*(\theta_k) \geq l^*(\theta^*) - c_k.$$

By compactness, there is a convergent subsequence of $\{\theta_k\}$ with limit θ_ϵ and

$$l^*(\theta_\epsilon) \geq l^*(\theta^*).$$

This contradicts the fact that θ^* is the unique maximiser, since $d(\theta_\epsilon, \theta^*) \geq \epsilon > 0$. (Note that we need compactness of the entire space here otherwise θ_k can go off to infinity.) □

Example 4.8 (conditions for Theorem 4.7 to hold for natural exponential families). Given $Y_1, \ldots, Y_n \overset{iid}{\sim} f_\theta$ with

$$f_\theta(y) = \exp\left(\theta y - A(\theta)\right) h(y) \in \mathcal{M} = \{f_\theta : \theta \in \mathcal{H}\},$$

what do we need so that the assumptions of Theorem 4.7 are satisfied?

1. $\mathcal{H}$ must be compact.
2. (a) Common support is always satisfied in exponential families.
   (b) $h(y)$ is measurable.
   (c) $A(\theta)$ is continuous.
3. We can bound the domination term, using the triangle inequality

   $$\int \sup_{\theta \in \mathcal{H}} |\theta y - A(\theta) + \log h(y)| f_{\theta^*}(y) dy \leq \sup_{\theta \in \mathcal{H}} |\theta| \mathbb{E}_{\theta^*}|Y| + \sup_{\theta \in \mathcal{H}} |A(\theta)| + \mathbb{E}_{\theta^*}|\log h(Y)|.$$

   (Here θ^* is the true data generating parameter.) $\sup_{\theta \in \mathcal{H}}|A(\theta)|$ is bounded since $\mathcal{H}$ is compact and A continuous. $\sup_{\theta \in \mathcal{H}}|\theta|$ is bounded since $\mathcal{H}$ is compact. Hence to ensure the domination condition is satisfied, we need only require that $\mathbb{E}_{\theta^*}(Y)$ and $\mathbb{E}_{\theta^*}|\log h(Y)|$ is bounded. (Since we generally do not know the true θ^*, we would have to require that these terms are bounded for all $\theta \in \mathcal{H}$ – but this bound doesn't have to be uniform!)

4.3 Generalisations of Theorem 4.7 to M- and Z-estimators

Using Proposition 4.4, Theorem 4.7 can be generalised to M- and Z-estimators. Further, Theorem 3.2 applies to M- and Z-estimators, as we spell out in the following theorem.

Theorem 4.9 (Theorem 3.2 generalised to M-estimators)**.** *Consider a sequence of (random) criterion functions* $\theta \mapsto \frac{1}{n}M_n(\theta)$ *and a function* $\theta \mapsto M^*(\theta)$ *such that*

1. *(Uniform convergence:)* $\sup_{\theta \in \mathcal{H}} \left| \frac{1}{n}M_n(\theta) - M^*(\theta) \right| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0$.
2. *(Well separated mode:)* θ^* *is the well separated mode of* M^**: for all* $\epsilon > 0$,

$$\sup_{\substack{\theta \in \mathcal{H} \\ d(\theta,\theta^*) \geq \epsilon}} M^*(\theta) < M^*(\theta^*).$$

If $\hat{\theta}_n$ *is a sequence of approximate maxima (or for Z-estimators, look at approximate roots of criterion functions), that is, if*

$$\frac{1}{n}M_n(\hat{\theta}_n) \geq \frac{1}{n}M_n(\theta) + o_{\mathbb{P}_*}(1),$$

then $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}_*} \theta^*$.

Theorem 4.10 (add-on Theorem 4.7 generalised to M-estimators)**.** *Let* $\{p_\theta : \theta \in \mathcal{H}\}$ *be a family of distributions and* $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(Y_i)$ *be a criterion function. Suppose we are given observations* $Y_1, \dots, Y_n \overset{iid}{\sim} p^*$. *Assume that*

1. $\mathcal{H}$ *is compact.*
2. (a) *The function* $y \mapsto m_\theta(y)$ *is measurable for all* $\theta \in \mathcal{H}$.

   (b) *The function* $\theta \mapsto m_\theta(y)$ *is continuous for all* $y \in \mathcal{Y}$.
3. *A domination condition:*

$$\int \sup_{\theta \in \mathcal{H}} |m_\theta(y)| p^*(dy) < \infty.$$

4. *The function*
$$\theta \mapsto M^*(\theta) = \mathbb{E}_*\left(m_\theta(Y)\right) = \int m_\theta(y)p^*(dy)$$
is uniquely maximised at θ^*.

5. $\hat{\theta}_n$ *is a maximiser of* $M_n(\theta)$.

Then the M-estimator is consistent: $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}_*} \theta^*$.

add-on We can weaken the assumption that $\mathcal{H}$ is compact – we need only the existence of a compact set $K \subset \mathcal{H}$ such that

$$\mathbb{E}_*\left[\sup_{\theta\in\mathcal{H}\cap K^c} m_\theta(Y)\right] < M^*(\theta^*).$$

(See Assignment 1, Question 5 for details and proof.)

add-on Consistency of Z-estimators is somewhat easier to establish. See Theorems 6.9.2 (page 513) of [LC06]. We also have an analogue to Theorem 3.2 (which uses stronger assumptions than Theorem 6.9.2 in [LC06]).

Theorem 4.11 (add-on Theorem 3.2 for Z-estimators)**.** *Let* Ψ_n *be random real-valued functions on* $\mathcal{H}$ *with roots* $\hat{\theta}_n$ *(not necessarily unique) and let* Ψ^* *be a fixed real-valued function with a unique root* θ^*. *(Note that uniqueness is implied by assumption 2.) Suppose*

1. *Uniform convergence:* $\sup_{\theta\in\mathcal{H}}|\Psi_n(\theta) - \Psi^*(\theta)| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0$.

2. *Well-separated root: Suppose that the root of* Ψ^* *is well separated: for all* $\epsilon > 0$,
$$\inf_{\substack{\theta\in\mathcal{H}\\ d(\theta,\theta^*)\geq\epsilon}} |\Psi^*(\theta)| > 0,$$
assuming that $\{\theta : d(\theta,\theta^*) \geq \epsilon\}$ *is non-empty.*

Then $\hat{\theta}_n$ *converges to* θ^* *in* $\mathbb{P}_*$*-probability. (Here* $\mathbb{P}_*$*-probability is the true generating process of the random variables* $\Psi_n(\theta)$*.)*

Proof. Establish $\Psi^\star(\hat{\theta}_n) \xrightarrow[n\to\infty]{\mathbb{P}_*} \Psi^\star(\theta^\star) = 0$:

$$\begin{aligned}\left|\Psi^\star(\hat{\theta}_n)\right| &= \left|\Psi^\star(\hat{\theta}_n) - \Psi^\star(\theta^\star)\right| \\ &= \left|\Psi^\star(\hat{\theta}_n) - \Psi_n(\hat{\theta}_n)\right| \\ &\leq \sup_{\theta\in\mathcal{H}}|\Psi^\star(\theta) - \Psi_n(\theta)| \\ &\xrightarrow[n\to\infty]{\mathbb{P}_*} 0,\end{aligned}$$

where the second line follows from the fact $\Psi^\star(\theta^\star) = 0 = \Psi_n(\hat{\theta}_n)$. That $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}_*} \theta^*$ follows exactly the second half of the proof of Theorem 3.2: Fix $\epsilon > 0$ and define

$$\eta(\epsilon) = \inf_{\substack{\theta\in\mathbb{R} \\ |\theta-\theta^*|\geq\epsilon}} |\Psi^*(\theta)|.$$

Then

$$\mathbb{P}_\star\left(\left|\hat{\theta}_n - \theta^\star\right| \geq \epsilon\right) \leq \mathbb{P}_\star\left(\left|\Psi^\star(\hat{\theta}_n)\right| \geq \eta(\epsilon)\right) \xrightarrow[n\to\infty]{\mathbb{P}_*} 0. \qquad \square$$

4.3.1 Weakening compactness

Compactness is a strong condition. The MLE can still be consistent when $\mathcal{H}$ is not compact. For example, if $\hat{\theta}^{\mathrm{ML}}$ is going to lie within a compact subset of $\mathcal{H}$ with overwhelming probability, then $\hat{\theta}^{\mathrm{ML}}$ will be consistent. (This can be formalised.)

add-on To prove consistency of an M-estimator, we can replace the compactness assumption with an assumption that there exists a compact set $K \subset \mathcal{H}$ such that $\theta^* \in K$ and

$$\mathbb{E}_*\left[\sup_{\theta\in\mathcal{H}\cap\mathcal{K}^c} m_\theta(Y)\right] < \mathbb{E}_*\left[m_{\theta^*}(Y)\right],$$

where θ^* is the maximiser of $\mathbb{E}_*\left[m_\theta(Y)\right]$ and the expectation is with respect to some p^*. With this assumption, we can show that with $\mathbb{P}_*$-probability going to one, the maximiser $\hat{\theta}_n$ of $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m(\theta)$ is in the compact set K. (See Assignment 1, Question 5.) Then apply the Theorem which assumes compactness to get consistency of $\hat{\theta}_n$.

4.3.2 Weakening uniform convergence

We can replace uniform convergence with weaker notions and get consistency of other estimators. One notion is epiconvergence, defined by $f_n \overset{e}{\to} f$ if , for all x and all sequences $\{x_n\}$ converging to x,

$$\liminf_{n\to\infty} f_n(x_n) \geq f(x),$$

and for all x and some sequence $x_n \to x$,

$$\limsup_{n\to\infty} f_n(x_n) \leq f(x).$$

Epiconvergence is used to get consistency of minimum distance estimators (MDE):

$$\hat{\theta}^{\mathrm{MDE}} := \operatorname*{argmin}_{\theta} d\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}, p_\theta\right),$$

where $\frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$ is the empirical PDF, and

$$\delta_y : t \mapsto \begin{cases} 1 & \text{if } t = y, \\ 0 & \text{otherwise.} \end{cases}.$$

These are similar to MoM estimators.

5 Lecture 9/2

5.1 Asymptotic distributional properties of the MLE

In the previous lectures we have established consistency of the MLE: $\hat{\theta}_n^{\mathrm{ML}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$. But we might want to ask how fast $\hat{\theta}_n^{\mathrm{ML}}$ converges to θ^* – or equivalently, how precise is $\hat{\theta}_n^{\mathrm{ML}}$ as an estimator of θ^*? That is, what is the sequence $\{r_n\}$ (the rate of convergence) so that $r_n(\hat{\theta}_n^{\mathrm{ML}} - \theta^*) \overset{d}{\to} F$ for some distribution F.

Recall why we study asymptotics. On the one hand, asymptotics is an abstraction, since we will always have a finite sample. But through asymptotics, we can unify many different parametric models. Moreover, asymptotics are good approximations if the rate of convergence r_n is good (i.e. $r_n = \Omega(\sqrt{n})$). To fully take advantage of asymptotics, it is therefore important to look beyond consistency and study the asymptotic distributional properties.

Recall from section 2.2.2, that for MLEs in full exponential families, we already have established an asymptotic distributional result:

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta^*)^{-1}\right),$$

in the well-specified setting (since in this setting the MLE can be expressed as a MoM estimator).

5.1.1 Score and Fisher information

Definition 5.1. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} p^*$ and $\mathcal{M} = \{p_\theta : \theta \in \mathcal{H}\}$. The *score statistic* is the first derivative of the log-likelihood (appropriately normalised):

$$S(\theta) = \frac{1}{n}\nabla_\theta l_n = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \log p_\theta(Y_i).$$

The *observed Fisher information* is the negative derivative of the score statistic:

$$-\frac{1}{n}\nabla_\theta^2 l_n = -\frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 \log p_\theta(Y_i),$$

where ∇_θ^2 is the Hessian.

(add-on I believe it is more appropriate to define the observed Fisher information as the square of the score statistic $S(\theta)^2$, since this always exists, whereas the log-likelihood may not be twice differentiable.)

The *expected Fisher information* is the expectation of the observed Fisher information (where the expectation is taken with respect to the same model parameter):

$$I(\theta) = \mathbb{E}_\theta\left[-\frac{1}{n}\nabla_\theta^2 l_n\right].$$

Remark 5.2. Suppose $Y_1, \ldots, Y_n \overset{iid}{\sim} p_\theta$. Then $\mathbb{E}_\theta S_\theta = 0$ assuming that you can swap derivatives and integrals. Moreover, $I(\theta) = \mathrm{Var}_\theta[S(\theta)] = \mathbb{E}_\theta[S(\theta)^2]$. (See Stat211 notes for proofs.)

5.1.2 Non-rigorous proof of MLE asymptotic normality

We can prove asymptotic normality of the MLE in more general settings than full natural exponential families. The gist of the proof follows.

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} p^*$ (so that we are not necessarily in the well-specified setting). Assuming consistency of the MLE, the fact that the MLE is a root of the score equation (i.e. it satisfies $\nabla_\theta l_n|_{\theta=\hat{\theta}_n^{\mathrm{ML}}} = 0$) suggests a Taylor expansion around the true parameter θ^*:

$$0 = \nabla_\theta l_n|_{\theta=\hat{\theta}_n^{\mathrm{ML}}} = \nabla_\theta l_n|_{\theta=\theta^*} + \left(\hat{\theta}_n^{\mathrm{ML}} - \theta^*\right) \nabla_\theta^2 l_n\big|_{\theta=\tilde{\theta}_n},$$

where $\tilde{\theta}_n$ is between θ^* and $\hat{\theta}_n^{\mathrm{ML}}$. Assuming $\nabla_\theta^2 l_n|_{\theta=\tilde{\theta}_n}$ is invertible,

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}} - \theta^*\right) = \left[-\frac{1}{\sqrt{n}}\, \nabla_\theta^2 l_n\big|_{\theta=\tilde{\theta}_n}\right]^{-1} \left[\frac{1}{\sqrt{n}}\, \nabla_\theta l_n|_{\theta=\theta^*}\right]. \tag{10}$$

For the second term, suppose that θ^* maximises $\mathbb{E}_* \log p_\theta(Y)$ so that

$$\mathbb{E}_*\left[\nabla_\theta \log p_\theta(Y)|_{\theta=\theta_*}\right] = 0,$$

(assuming EDI). Then we know by the CLT that

$$\begin{aligned}\frac{1}{\sqrt{n}}\, \nabla_\theta l_n|_{\theta=\theta^*} &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log p_\theta(Y_i)|_{\theta=\theta^*} - \mathbb{E}_*\left[\nabla_\theta \log p_\theta(Y)|_{\theta=\theta^*}\right]\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}_*\left[\nabla_\theta \log p_\theta(Y)|_{\theta=\theta^*}\right]\right).\end{aligned} \tag{11}$$

For the first term, we assumed consistency, so we get $\tilde{\theta}_n \xrightarrow{\mathbb{P}_*} \theta^*$ and we would expect

$$\left[\frac{1}{n}\, \nabla_\theta^2 l_n\big|_{\theta=\tilde{\theta}_n}\right]^{-1} \xrightarrow{\text{CMT}} \left[\frac{1}{n}\, \nabla_\theta^2 l_n\big|_{\theta=\theta^*}\right]^{-1} \xrightarrow{\text{LLN}} \left[\mathbb{E}_\star\, \nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^*}\right]^{-1}. \tag{12}$$

Combining (10), (11), (12), we get

$$\begin{aligned}&\sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}} - \theta^*\right) \xrightarrow{d} \\ &\quad \mathcal{N}\left(0, \left[\mathbb{E}_\star\, \nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^*}\right]^{-1} \mathrm{Var}_*\left[\nabla_\theta \log p_\theta(Y)|_{\theta=\theta^*}\right] \left[\mathbb{E}_\star\, \nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^*}\right]^{-1}\right)\end{aligned} \tag{13}$$

(The Hessian is symmetric, so no need to include a transpose in the last term of the variance.) This is called the sandwich formula for variance. In the well-specified case

$$I(\theta^*) = -\mathbb{E}_\star \nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^*} = \mathrm{Var}_* \left[\nabla_\theta \log p_\theta(Y)|_{\theta=\theta^*}\right],$$

as long as EDI holds and so (13) simplifies to

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta^*)^{-1}\right).$$

In the misspecified case, we can't simplify (13) further.

5.1.3 Sufficient conditions for exchanging derivatives and integrals (EDI)

Proposition 5.3. *Suppose that*

1. *V is an open subset of $\mathbb{R}^p$;*
2. *$(S, \mathcal{A}, \mu)$ is a measure space;*
3. *$f : V \times S \to \mathbb{R}$ is μ-integrable for every $v \in V$;*
4. *for every $(v, s) \in V \times S$, the derivative $\frac{\partial}{\partial v} f(v, s)$ exists;*
5. *$v \mapsto \frac{\partial}{\partial v} f(v, s)$ is continuous for every $s \in S$;*
6. *Suppose there exists a μ-integrable function $g : S \to \mathbb{R}$ such that for all $v \in V, s \in S$,*
$$\left\|\frac{\partial}{\partial v} f(v, s)\right\| \leq g(s).$$

Then $\phi : v \mapsto \int_S f(v, s) d\mu(s)$ is differentiable and

$$\frac{\partial}{\partial v}\phi(v) = \int_S \frac{\partial}{\partial v} f(v, s) d\mu(s).$$

Remark 5.4. In the proof of $\mathbb{E}_\theta S(\theta) = 0$, we used the above proposition with $f(\theta, y) = p_\theta(y)$ and $g(y) = \sup_{\theta\in\mathcal{H}} \|\nabla_\theta p_\theta(y)\|$. In the proof of $I(\theta) = \mathrm{Var}_\theta[S(\theta)]$, we used the proposition with $f(\theta, y) = \nabla_\theta p_\theta(y)$ and $g(y) = \sup_{\theta\in\mathcal{H}} \|\nabla_\theta^2 p_\theta(y)\|$.

5.1.4 Asymptotic normality of the MLE

This section makes rigorous the discussion from section 5.1.2 and collects all of the assumptions into a precise statement

Theorem 5.5. *Let $\mathcal{M} = \{p_\theta : \theta \in \mathcal{H}\}$ be our model and assume that the p_θ have common support $\mathcal{Y}$. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} p^*$. Suppose that*

1. *θ^* is a local maximiser of $l^*(\theta) = \mathbb{E}_*\left[\log p_\theta(Y)\right]$.*
2. *The MLE is consistent: $\hat{\theta}_n^{ML} \xrightarrow[n\to\infty]{\mathbb{P}_*} \theta^*$ (where the MLE is any root of the score equation).*
3. *There exists an open set U such that*
   - *(a) $\theta^* \in U$;*
   - *(b) for all $y \in \mathcal{Y}$, the map $\theta \mapsto p_\theta(y)$ is twice continuously differentiable on U;*
   - *(c) $y \mapsto p_\theta(y)$ is measurable for all $\theta \in U$.*
4. *$\mathbb{E}_*\left[\nabla_\theta^2 \log p_\theta(Y)|_{\theta=\theta^*}\right]$ is non-singular and $\mathbb{E}_* \| \nabla_\theta^2 \log p_\theta(Y)|_{\theta=\theta^*} \|^2 < \infty$. (The norm $\|\cdot\|$ is the sum of squared entries of the Hessian.)*
5. *A domination condition: There exists a compact ball $K \subset U$ centred at θ^* such that*
   - *(a) $\mathbb{E}_*\left[\sup_{\theta \in K} \|\nabla_\theta \log p_\theta(Y)\|\right] < \infty$,*
   - *(b) $\mathbb{E}_*\left[\sup_{\theta \in K} \|\nabla_\theta^2 \log p_\theta(Y)\|\right] < \infty$.*

   (The norm $\|\cdot\|$ is the sum of absolute values of entries of the gradient/Hessian.)

Then as $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_n^{ML} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \left[\mathbb{E}_\star\, \nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^*}\right]^{-1} \mathrm{Var}_*\left[\nabla_\theta \log p_\theta(Y)|_{\theta=\theta^*}\right] \left[\mathbb{E}_\star\, \nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^*}\right]^{-1}\right) \tag{14}$$

add-on The maximiser θ^* does not have to be unique, nor does it even need to be a global – a local maximum works fine. But typically θ^* will be a unique global maximiser: In the well-specified case, Proposition 3.1 showed that the maximiser is unique under mild conditions. In the misspecified case, $l^*(\theta)$ is maximised at the KL-projection of p^* onto $\mathcal{M}$; if $\mathcal{M}$ is closed, non-empty and convex then the KL-projection is unique.

See Theorems 5.41 and 5.42 from [vdV] for a slightly more general result without the ambiguity in the definitions of θ^* and $\hat{\theta}_n^{\mathrm{ML}}$.

The proof of Theorem 5.5 will be presented in the next lecture.

Remark 5.6. Consider the family of Laplace distributions with scale 1 and mean $\theta \in \mathbb{R}$:

$$p_\theta(y) = \frac{1}{2} \exp\left(-|y - \theta|\right).$$

The MLE for θ is the median. p_θ does not satisfy assumption 3b of Theorem 2.7, yet the MLE is asymptotically normal. This example illustrates how weak the theorem is.

Theorem 5.5 is a classical result. We will see more general and powerful results in the following lectures.

6 Lecture 11/2

6.1 Proof of Theorem 5.5 (classical MLE asymptotics)

Proof of Theorem 5.5. It suffices to show convergence in distribution only on sets $\{A_n\}$ with $\mathbb{P}(A_n) \to 1$. That is, we will show that

$$\mathbb{P}(X_n \leq x \cap A_n) \to F(x),$$

pointwise, at all continuity points x of the CDF F of (14). [This is a useful trick!]

The MLE $\hat{\theta}_n^{\mathrm{ML}}$ has the property that $\frac{1}{n}\nabla_\theta l_n(\hat{\theta}_n^{\mathrm{ML}}) = 0$. So the mean value theorem (i.e. first-order Taylor expansion) says

$$0 = \frac{1}{n}\nabla_\theta l_n(\theta^*) + \frac{1}{n}\nabla_\theta^2 l_n(\tilde{\theta}_n)(\hat{\theta}_n^{\mathrm{ML}} - \theta^*), \tag{15}$$

where $\tilde{\theta}_n$ is between $\hat{\theta}_n^{\mathrm{ML}}$ and θ^*. But we already know from the CLT

$$\sqrt{n}\left(\frac{1}{n}l_n(\theta^*) - \mathbb{E}_*\left[\nabla_\theta \log p_{\theta^*}(Y)\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}_*\nabla_\theta \log p_{\theta^*}(Y)\right).$$

Since θ^* is a local maximiser of $\mathbb{E}_* \log p_\theta(Y)$, the second term on the LHS of the above equation is zero by EDI. (add-on I think this is why we need assumption 5.(a).)

Our focus is therefore on the second term in (15). We will prove

$$\frac{1}{n}\nabla_\theta^2 l_n(\tilde{\theta}_n) \xrightarrow{\mathbb{P}_*} \mathbb{E}_*\nabla_\theta^2 \log p_{\theta^*}(Y), \tag{16}$$

at least on a set A_n with probability going to 1. Once we've established the LHS of (16) is non-singular (again, at least on a set B_n with probability going to 1), a simple application of Slutsky's theorem would complete the proof.

To prove (16), we start by noting $\tilde{\theta}_n$ is consistent, since $\hat{\theta}_n^{\mathrm{ML}}$ is:

$$\mathbb{P}_*\left(d(\tilde{\theta}_n, \theta^*) > \epsilon\right) \le \mathbb{P}_*\left(d(\hat{\theta}_n^{\mathrm{ML}}, \theta^*) > \epsilon\right) \xrightarrow{n\to\infty} 0.$$

Secondly,

$$\begin{aligned}\left|\frac{1}{n}\nabla_\theta^2 l_n(\tilde{\theta}_n) - \mathbb{E}_*\nabla_\theta^2 \log p_{\theta^*}(Y)\right| &\le \left|\frac{1}{n}\nabla_\theta^2 l_n(\tilde{\theta}_n) - \mathbb{E}_*\nabla_\theta^2 l_n(\tilde{\theta}_n)\right| \\ &\quad + \left|\mathbb{E}_*\nabla_\theta^2 l_n(\tilde{\theta}_n) - \mathbb{E}_*\nabla_\theta^2 \log p_{\theta^*}(Y)\right| \\ &\le \sup_{\theta\in K}\left|\frac{1}{n}\nabla_\theta^2 l_n(\theta_n) - \mathbb{E}_*\nabla_\theta^2 l_n(\theta_n)\right| \\ &\quad + \left|\mathbb{E}_*\nabla_\theta^2 l_n(\tilde{\theta}_n) - \mathbb{E}_*\nabla_\theta^2 \log p_{\theta^*}(Y)\right|, \end{aligned} \tag{17}$$

where the second line holds when $\tilde{\theta}_n \in K$, which is true with probability going to 1. Applying ULLN (Proposition 4.4, whose assumptions hold by 5.(b)), we get that

$$\sup_{\theta\in K}\left|\frac{1}{n}\nabla_\theta^2 l_n(\theta_n) - \mathbb{E}_*\nabla_\theta^2 l_n(\theta_n)\right| \xrightarrow[n\to\infty]{\mathbb{P}_*} 0.$$

Use DCT to show that the second term in (17) converges to zero. (In order to apply the DCT, we need the domination condition 5.(b) and we need to show that $\theta \mapsto \nabla_\theta^2 \log p_\theta(y)$ is continuous, so that $\nabla_\theta^2 l_n(\tilde{\theta}_n) \to \nabla_\theta^2 \log p_{\theta^*}(Y)$. We did a similar proof of continuity last week.) This proves (16).

Now

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) = \left[-\frac{1}{n}\nabla^2_\theta l_n(\tilde{\theta}_n)\right]^{-1}\left(\frac{1}{\sqrt{n}}\nabla_\theta l_n(\theta^*)\right),$$

only if the Hessian $\nabla^2_\theta l_n(\tilde{\theta}_n)$ is non-singular. We know that $\mathbb{E}_*\left[\nabla^2_\theta \log p_\theta(Y)|_{\theta=\theta^*}\right]$ is non-singular and that the Hessian converges in probability to this expectation. It is left as an exercise to show that there exists events B_n with probability going to 1 such that the Hessian is non-singular on B_n. (add-on Hint: use the assumptions that the likelihood is twice continuously differentiable on U, so that non-singularity extends to an open ball around the expectation; and then observe that the Hessian must be in this open ball with probability going to 1.) □

The proof only holds for univariate θ (but the Theorem holds more generally). We need to change the Taylor series expansion to allow for multivariate θ. Here $\nabla^2_\theta l_n(\tilde{\theta})$ becomes the matrix of 2nd derivatives of l_n with the j-th row evaluated at some $\tilde{\theta}_{n,j}$ between θ^* and $\hat{\theta}_n$. Then change

$$\frac{1}{n}\lambda^n_\theta l_n(\tilde{\theta}_n) \xrightarrow{\mathbb{P}_*} \mathbb{E}_*\nabla^2_\theta \log p_{\theta^*}(Y)$$

to

$$\frac{1}{n}\frac{\partial^2}{\partial\theta_j\partial\theta_k} l_n(\tilde{\theta}_{n,j}) \xrightarrow{\mathbb{P}_*} \mathbb{E}_*\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_{\theta^*}(Y),$$

for all j, k. See [FWCT] for details on the multivariate Taylor series expansion.

6.2 Quadratic mean differentiability (QMD)

We saw in Example 5.6 that Theorem 5.5 relies on strong assumptions which often don't hold, even when the MLE is asymptotically Gaussian. This motivates the development of more general smoothness conditions which still guarantee that the MLE is asymptotically Gaussian. This smoothness condition – called quadratic mean differentiability (QMD) – will replace the twice continuously differentiable likelihood assumption in Theorem 5.5.

QMD is a regularity condition on the map $\theta \mapsto \sqrt{p_\theta}$ rather than on the map $\theta \mapsto p_\theta$. We will see why this is relevant and powerful.

Definition 6.1. A family $\{p_\theta : \theta \in \mathcal{H}\}$ with $\mathcal{H} \subset \mathbb{R}^p$ is *quadratically mean differentiable* at θ_0 if there exists a vector-valued function $\eta(x, \theta_0)$ such that for all $h \in \mathbb{R}^p$,

$$\int_{\mathcal{X}} \left(\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{\eta(x,\theta_0)^\intercal h}{2} \sqrt{p_{\theta_0}(x)} \right)^2 d\mu(x) = o(\|h\|^2) \text{ as } h \to 0. \quad (18)$$

The function η is called the *quadratic mean derivative* (or sometimes the *score*).

Intuition: The integrand in (18) looks like a first-order Taylor expansion. If $\sqrt{p_\theta}$ is twice continuously differentiable at θ_0, then

$$\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \left(\nabla_\theta \sqrt{p_\theta(x)}\Big|_{\theta=\theta_0} \right)^\intercal h = \frac{1}{2}\, \nabla^2_\theta \sqrt{p_\theta(x)}\Big|_{\theta=\theta_0} h^2 = o(\|h\|).$$

In this case we would expect

$$\frac{\eta(x,\theta_0)^\intercal h}{2} \sqrt{p_{\theta_0}(x)} = \left(\nabla_\theta \sqrt{p_\theta(x)}\Big|_{\theta=\theta_0} \right)^\intercal h,$$

so that

$$\frac{\eta(x,\theta_0)}{2} \sqrt{p_{\theta_0}(x)} = \nabla_\theta \sqrt{p_\theta(x)}\Big|_{\theta=\theta_0} = \frac{\frac{\partial}{\partial\theta} p_\theta(x)\big|_{\theta=\theta_0}}{2\sqrt{p_{\theta_0}(x)}},$$

and hence

$$\eta(x,\theta_0) = \frac{\frac{\partial}{\partial\theta} p_\theta(x)\big|_{\theta=\theta_0}}{p_\theta(x)} = \frac{\partial}{\partial\theta} \log p_\theta(x)\Big|_{\theta=\theta_0}.$$

So intuitively η should be viewed as the 'derivative' of the log-likelihood – justifying it's name as the score function.

Remark 6.2.

1. Why is QMD better than assuming differentiability of $\theta \mapsto \log p_\theta(x)$? There are some functions which are QMD but not differentiable (for example the double Laplace distribution from the previous section).

   Since the QMD takes an integral over $\mathcal{X}$, we are able to ignore sets of measure zero. In particular, QMD can ignore the fact that the derivative may not exist at some points (of measure zero).

2. Why do we define QMD using $\sqrt{p_\theta(x)}$ and not $p_\theta(x)$? This is so that we don't require L^2-integrability.

Example 6.3.

1. In section, we will prove that the double Laplace distribution
$$p_\theta(x) = \frac{1}{2}e^{-|x-\theta|},$$
is QMD. How do we guess what η should be? Where the derivative of the log-likelihood l exists, we can set η to equal l'. The points where l' doesn't exist form a measure-zero set, so η can be arbitrary on these points and the integral in (18) will remain unchanged. For example,
$$\eta(x,\theta_0) = \begin{cases} 1 & \text{if } x < \theta_0, \\ -1 & \text{if } x > \theta_0, \\ 0 & \text{if } x = \theta_0, \end{cases}$$
will satisfy the definition of QMD.

2. Let $p_\theta = \text{Unif}([0,\theta])$. Is this family QMD at any θ_0? The answer is no, because the support depends on θ. Compute
$$\begin{aligned} p_\theta(x) &= \frac{\mathbb{1}\{x \in [0,\theta]\}}{\theta} \\ \log p_\theta(x) &= \log \mathbb{1}\{\theta \in [x,\theta]\} - \log\theta. \end{aligned}$$
Take $\theta_0 > 0$ and $h > 0$. Then
$$\begin{aligned} &\int_{\theta_0}^{\theta_0+h} \left(\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2}\eta(x,\theta_0)h\sqrt{p_{\theta_0}(x)}\right)^2 d\mu(x) \\ &\qquad = \int_{\theta_0}^{\theta_0+h} \left(\frac{1}{\sqrt{\theta_0+h}}\right)^2 d\mu(x) \\ &\qquad = \frac{h}{\theta_0+h} \neq o(\|h\|^2), \end{aligned}$$
since $p_{\theta_0}(x) = 0$ for $x \in (\theta_0, \theta_0+h]$.

6.2.1 QMD generalises Fisher information

Previously, when we wanted to understand the MLE for exponential families, we used Fisher information. Now we want a generalisation of Fisher information in QMD families.

Theorem 6.4 (Lemma 12.2.1 of [TSH])**.** *If the parameter space $\mathcal{H}$ is an open subset of $\mathbb{R}^k$ and the model is QMD at $\theta_0 \in \mathcal{H}$, then*

1. $\mathbb{E}_{\theta_0}\left[\eta(X,\theta_0)\right] = 0$*, and*
2. $\mathbb{E}_{\theta_0}\left[\eta_i(X,\theta_0)\eta_j(X,\theta_0)\right] < \infty$*, so all the elements of* $\mathbb{E}_{\theta_0}\eta(X,\theta_0)\eta(X,\theta_0)^\mathsf{T}$ *are finite.*

See proof in Section 3.

Definition 6.5. Suppose p_θ is QMD at an interior point θ_0 of the parameter space $\mathcal{H}$. Then the Fisher information of θ_0 is given by

$$I(\theta_0) = \mathbb{E}_{\theta_0}\eta(X,\theta_0)\eta(X,\theta_0)^\mathsf{T},$$

which is well defined by the previous theorem.

When the likelihood $\theta \mapsto p_\theta(x)$ is continuously differentiable and I continuous, this definition agrees with the original definition of Fisher information:

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right]^\mathsf{T}\right),$$

by Theorem 6.6.

6.2.2 Sufficient conditions for QMD

Theorem 6.6 (Lemma 7.6 of [vdV])**.** *For every θ in an open subset Θ of $\mathbb{R}^k$, let p_θ be a μ-density. Suppose that the map*

$$\Theta \to \mathbb{R},$$

$$\theta \mapsto \sqrt{p_\theta(x)},$$

is continuously differentiable for all x. If the elements of the Fisher information matrix

$$I(\theta) := \int_{\mathcal{X}} \frac{\nabla_\theta p_\theta(x) \left(\nabla_\theta p_\theta(x)\right)^{\mathsf{T}}}{p_\theta(x) p_\theta(x)} p_\theta(x) d\mu(x),$$

are well defined and continuous in θ, for all $\theta \in \Theta$, then p_θ is QMD at every $\theta \in \Theta$ with score

$$\eta(X, \theta) = \frac{\frac{\partial p_\theta(X)}{\partial \theta}}{p_\theta(X)} = \frac{\partial}{\partial \theta} \log p_\theta(X).$$

The assumptions of this theorem can hold when p_θ is not twice continuously differentiable. See also theorem 3 in section 3, for another set of (slightly different) sufficient conditions for QMD.

6.3 General MLE asymptotic Normality theorem

Recall that we introduced QMD so that we could develop an asymptotic normality theorem for MLEs which didn't rely on twice continuous differentiability. We are now ready to state that theorem.

Theorem 6.7 (Theorem 7.12 of [vdV]). *Let $X_1, \ldots, X_n \overset{iid}{\sim} p_{\theta^*}$ with θ^* in the interior of the parameter space $\mathcal{H}$. Assume p_θ is QMD at θ^*. (A locally Lipschitz condition:) Suppose there exists a measurable function $K(x) \in L^2(\theta^*)$ (i.e. $\mathbb{E}_{\theta^*} K^2(X) < \infty$) such that*

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq K(x) \|\theta_1 - \theta_2\|,$$

for all x and all θ_1, θ_2 in a neighbourhood of θ^*. If $I(\theta^*)$ (as given in Definition 6.5) is non-singular and the MLE $\hat{\theta}_n^{ML}$ consistent for θ^*, then

$$\sqrt{n}\left(\hat{\theta}_n^{ML} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}(\theta^*)\right).$$

There is analogous theorem for the misspecified case (it just requires much more notation). This is much stronger than Theorem 5.5 – we have replaced five strong assumptions with a single smoothness condition (QMD) and a locally Lipschitz condition.

Intuition for the locally Lipschitz condition: Suppose $\log p_\theta(x)$ is differentiable with regard to θ. Then the MVT says

$$\log p_{\theta_1}(x) - \log p_{\theta_2}(x) = \frac{\partial}{\partial\theta} \log p_\theta(x)\bigg|_{\theta=\tilde{\theta}} (\theta_1 - \theta_2)\,.$$

So, if we can find a square-integrable function $K(x)$ that bounds $\frac{\partial}{\partial\theta}\log p_\theta(x)$ in a neighbourhood around θ^*, then K will satisfy the locally Lipschitz condition.

7 Lecture 16/2

In this lecture, we will extend the MLE asymptotic normality theorem to M- and Z-estimators and then examine some applications (asymptotic influence functions, robust regression and optimal robust estimators).

7.1 Recap of $o_{\mathbb{P}}$ and $O_{\mathbb{P}}$

Recall the definitions of stochastic o and O symbols from Definition 3.4:

1. $o_{\mathbb{P}}(1)$ is shorthand for a sequence of random variables that converge to zero in $\mathbb{P}$-probability.

2. $o_{\mathbb{P}}(R_n)$ is a sequence of random variables $X_n = R_n Y_n$ with $Y_n = o_{\mathbb{P}}(1)$.

3. $O_{\mathbb{P}}(1)$ is a sequence of random variables X_n that are uniformly tight (aka bounded in probability): for every $\epsilon > 0$, there exists M such that $\mathbb{P}(|X_n| > M) < \epsilon$ for all n.

4. $O_{\mathbb{P}}(R_n)$ is defined analogously to $o_{\mathbb{P}}(R_n)$.

When $R_n < 1$, R_n^{-1} can be understood as the 'rate' (i.e. speed) of convergence. For example, if $X_n = o_{\mathbb{P}}(1/\sqrt{n})$ then $X_n \xrightarrow{\mathbb{P}} 0$ faster than $1/\sqrt{n}$. When $R_n > 1$, it can be thought of as a dominator of X_n. If $X_n = o_{\mathbb{P}}(R_n)$, then R_n grows asymptotically a factor faster than X_n – so much so that X_n/R_n tends to zero in probability.

7.2 Asymptotic distributions of M- and Z-estimators

We can extend the MLE asymptotic results to M- and Z-estimators with some regularity conditions.

Theorem 7.1 (Z-estimators - Theorem 5.21 of [vdV])**.** *Let* $Y_1, \dots, Y_n \overset{iid}{\sim} p^*$. *For every* θ *in an open subset* $\mathcal{H}$ *of* $\mathbb{R}^p$, *let* ψ_θ *be a measurable vector-valued function (i.e. we have a function* $\mathcal{H} \times \mathcal{Y} \to \mathbb{R}_k$ *given by* $(\theta, y) \mapsto \psi_\theta(y)$*). Assume that*

1. *(A local Lipschitz condition:) There exists a measurable* L^2 *function* $\dot{\psi} : \mathcal{Y} \to \mathbb{R}$ *(i.e.* $\int_{\mathcal{Y}} \dot{\psi}^2(y) p^*(dy) < \infty$*) such that for all* θ_1, θ_2 *in a neighbourhood of* θ^*,

   $$\|\psi_{\theta_1}(y) - \psi_{\theta_2}(y)\| \leq \dot{\psi}(y) \|\theta_1 - \theta_2\|.$$

   ($\|\cdot\|$ *is the Euclidean norm.)*

2. $\theta \mapsto \int_{\mathcal{Y}} \psi_\theta(y) p^*(dy)$ *is differentiable at a zero* θ^* *with non-singular derivative matrix* V^*.

3. $\int_{\mathcal{Y}} \|\psi_{\theta^*}(y)\| p^*(dy) < \infty$.

4. $\frac{1}{n} \sum_{i=1}^n \psi_{\hat{\theta}_n}(Y_i) = o_{\mathbb{P}^*}(1/\sqrt{n})$. *(That is* $\hat{\theta}_n$ *is an approximate zero, or,* $\hat{\theta}_n$ *are approximate Z-estimators with respect to* ψ_θ*.)*

5. $\hat{\theta}_n \xrightarrow[n \to \infty]{\mathbb{P}_*} \theta^*$ *as* $n \to \infty$.

Then

$$\sqrt{n} \left(\hat{\theta}_n - \theta^* \right) = -{V^*}^{-1} \frac{1}{n} \sum_{i=1}^n \psi_{\theta^*}(Y_i) + o_{\mathbb{P}_*}(1).$$

In particular

$$\sqrt{n} \left(\hat{\theta}_n - \theta^* \right) \xrightarrow{d} \mathcal{N} \left(0, {V^*}^{-1} \mathbb{E}_* \left[\psi_{\theta^*}(Y) \psi_{\theta^*}(Y)^\mathsf{T} \right] {{V^*}^{-1}}^\mathsf{T} \right)$$

(add-on Note: $\mathbb{E}_* \left[\psi_{\theta^*}(Y) \psi_{\theta^*}(Y)^\mathsf{T} \right] = \mathrm{Cov}[\psi_{\theta^*}(Y)]$ since $\mathbb{E}_* \left[\psi_{\theta^*}(Y) \right] - 0$.)

To get the analogous theorem for M-estimators, roughly we will integrate ψ from the Z-estimator theorem to produce our M-estimator.

Theorem 7.2 (M-estimators - Theorem 5.23 of [vdV]). *Let* $Y_1, \dots, Y_n \overset{iid}{\sim} p^*$. *Let* m_θ *be a measurable scalar-valued function such that*

1. θ^* *is a maximum of* $\theta \mapsto \int_{\mathcal{Y}} m_\theta(y) p^*(dy)$;
2. $\theta \mapsto m_\theta(y)$ *is differentiable at* θ^* *for* $\mathbb{P}_*$*-almost every* y *(in fact it suffices that the map is differentiable at* θ^* *in* $\mathbb{P}_*$*-probability), with derivative* $m'_{\theta^*}(y)$;
3. *there exists a measurable function* $\dot{m} : \mathcal{Y} \to \mathbb{R}$ *in* L^2 *with*

   $$|m_{\theta_1}(y) - m_{\theta_2}(y)| \leq \dot{m}(y) \|\theta_1 - \theta_2\|,$$

   for all θ_1, θ_2 *in an open neighbourhood of* θ^*.
4. $\theta \mapsto \int_{\mathcal{Y}} m_\theta(y) p^*(dy)$ *admits a second-order Taylor expansion (i.e. it is twice differentiable) at* θ^*, *with non-singular second derivative matrix* V^*.
5. $\hat{\theta}_n$ *are approximate M-estimators:*

   $$\frac{1}{n} \sum_{i=1}^n m_{\hat{\theta}_n}(Y_i) \geq \sup_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n m_\theta(Y_i) - o_{\mathbb{P}_*}\left(\frac{1}{n}\right);$$

6. $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}_*} \theta^*$ *as* $n \to \infty$.

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) = -V^{*-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n m'_{\theta^*}(Y_i) + o_{\mathbb{P}_*}(1).$$

In particular,

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, V^{*-1} \mathbb{E}_*\left[m'_{\theta^*}(Y) m'_{\theta^*}(Y)^\mathsf{T}\right] V^{*-1}\right)$$

Example 7.3. Consider the criterion function

$$m_\theta(x) = (1-\alpha)[\theta - x]_+ + \alpha[x - \theta]_+,$$

where $\alpha \in (0,1)$. Suppose X is drawn from a continuous distribution. The α-th quantile

$$q_\alpha = \inf\{q \in R | \alpha \leq \mathbb{P}(X \leq q)\}$$

minimises $\mathbb{E}[m_\theta(X)]$ since

$$\frac{\partial}{\partial\theta}\mathbb{E}[m_\theta(X)] = (1-\alpha)\mathbb{P}(X \leq \theta) - \alpha\mathbb{P}(X \geq \theta) = \mathbb{P}(X \leq \theta) - \alpha,$$

(use EDI and the fact that the derivative doesn't exist only at $\theta = x$, which occurs with probability 0) and $\mathbb{P}(X \leq q_\alpha) = \alpha$ since X is continuous. We can check the M-estimator theorem conditions:

1. We can choose $\dot{m}(x) = 1$ since
$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq |\theta_1 - \theta_2|.$$

2. With probability 1, $X \neq \theta$, in which case the derivative exists:
$$m'_\theta(X) = (1-\alpha)\mathbb{1}\{\theta \geq X\} - \alpha\mathbb{1}\{\theta \leq X\}\,.$$

3. Twice differentiability of $\mathbb{E}[m_\theta(X)]$ holds:
$$\frac{\partial}{\partial\theta}\mathbb{E}[m_\theta(X)] = (1-\alpha)\mathbb{P}(X \leq \theta) - \alpha\mathbb{P}(X \geq \theta) = \mathbb{P}(X \leq \theta) - \alpha,$$
$$\frac{\partial^2}{\partial\theta^2}\mathbb{E}[m_\theta(X)] = f(\theta).$$
Further, $f(q_\alpha) > 0$, so the non-singularity condition is satisfied.

4. Exercise: find $\hat{\theta}_n$ and check consistency.

7.3 Asymptotic influence functions (AIF) and robustness

We want to answer the question: "what is the influence of a single data point on the estimator, as $n \to \infty$?" Recall the conclusion of Theorem 7.2:

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) = -V^{*-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m'_{\theta^*}(Y_i) + o_{\mathbb{P}_*}(1).$$

Then the difference in the estimator caused by the n-th sample Y_n is given by:

$$\hat{\theta}_n - \hat{\theta}_{n-1} = -\frac{1}{n}V^{*-1}m'_{\theta^*}(Y_n) + o_{\mathbb{P}_*}(1). \tag{19}$$

(The difference between $\frac{1}{n}$ and $\frac{1}{n-1}$ can be absorbed in the $o_{\mathbb{P}_*}(1)$ term.)

Definition 7.4. For an M-estimator with criterion function m_θ,

$$y \mapsto V^{*-1} m'_{\theta^*}(y),$$

is called the *asymptotic influence function* (AIF). For a Z-estimator with estimating equations $\sum_{i=1}^n \psi(Y_i) = 0$, the asymptotic influence function is given by

$$y \mapsto V^{*-1} \psi(y).$$

(This definition is in the context of the regularity assumptions of Theorems 7.1 and 7.2.)

An estimator is called robust if it is influenced too much by outliers or extreme values. (This is deliberately left as a loose notion.) To ensure robustness of M- and Z-estimators, it suffices that $m'_{\theta^*}(\cdot)$ is bounded since this will ensure (19) goes to zero. This is called B-robustness.

Definition 7.5. An M-estimator is *B-robust* if its AIF is bounded.

7.3.1 Detour: AIF in high dimensions

In high dimensions, we don't have consistency, so classical asymptotic distribution results don't apply. But it turns out that we can recover much of the classical theory through AIFs.

The canonical example is logistic regression:

$$Y_i \overset{iid}{\sim} \mathrm{Bern}(\sigma(X_i^\mathsf{T} \beta_0)),$$

where $\sigma(x) = \frac{e^x}{1+e^x}$ is the sigmoid function. In the high dimensional setting, the dimension $p(n)$ of β_0 can vary as a function of n: it is assumed that $\frac{p(n)}{n} \to k > 0$. The MLE is no longer consistent, but all hope is not lost since we know that the MLE is a root of the score function. With some working, we can use this fact to find the influence function. And we can leverage the AIF to build parallel asymptotic theory in high dimensions. This is called the leave-one-out trick in statistics (see [EKBB$^+$13, SC18]); in probability theory and statistical physicals, this is called the cavity method (and is an important technique for random matrix theory). It is also used in ML (see [CLC19, HL20]).

7.3.2 Robust regression

(Example 5.28 of [vdV]:) Consider the following setup: $(X_1, Y_1), \dots, (X_n, Y_n)$ are iid with $Y_i = X_i^\mathsf{T}\theta + e_i$ with $e_i \overset{iid}{\sim} F_e$ independent of the X_i's. We can express squared error loss (not robust) and absolute loss estimators as M-estimators with

$$M(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(Y_i - \theta^\mathsf{T} X_i).$$

(We have changed notation slightly: the m_θ in Theorem (7.2) is now $m(y - \theta^\mathsf{T} x)$.) If we want B-robustness, we need to bound the derivative $(y, x) \mapsto m'(y - \theta^\mathsf{T} x_i)x$. If X_i may be unbounded, then bounding m' is not enough to ensure B-robustness!

This discussion provides a guide to developing robust estimators: 1) examine an existing M-estimator based on the criterion function m_θ; 2) tweak m'_θ to some $\tilde{m}'_\theta$ which is bounded; 3) use the tweaked $\tilde{m}'_\theta$ as the criterion function of a Z-estimator. For example, the above regression estimator looks like

$$\sum_{i=1}^{n} \psi(Y_i - \theta^\mathsf{T} X_i)X_i = 0, \tag{20}$$

as a Z-estimator. How do we make this robust? One idea is to replace the estimating equation (20) with

$$\sum_{i=1}^{n} \psi(Y_i - \theta^\mathsf{T} X_i)X_i = 0,$$

where ψ and ν are bounded. What are the optimal choices of ψ and ν?

7.3.3 Optimal robust location estimators

(Example 5.29 of [vdV]:)Loosely, an robust estimator is optimal if it has the minimum variance while maintaining a certain degree of robustness. For M- and Z-estimators, we can formalise the robustness requirement as thresholding the AIF; then our problem becomes one of constrained optimisation: minimise variance subject to a bound on the AIF.

We will focus on location estimators in the context of $X_1, \ldots, X_n \overset{iid}{\sim} p_*$. Any function ψ defines a location estimator $\hat{\theta}_n$ as *a* solution to

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0.$$

For this section, assume whatever conditions are needed on ψ to guarantee the assumptions of Theorem (7.1) hold. Assume that $\hat{\theta}_n$ is consistent for the unique solution $\theta_0 = 0$ to

$$\mathbb{E}_* \left[\psi(X - \theta)\right] = 0. \tag{21}$$

Suppose $\psi : \mathbb{R} \to \mathbb{R}$ and assume EDI so that the derivative V^* to $\theta \mapsto \mathbb{E}_* \left(\psi(X - \theta)\right]$ at $\theta_0 = 0$ is $-\mathbb{E}_* \psi'(X)$. Then Theorem (7.1) gives

$$\sqrt{n}\hat{\theta}_n \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}_* \psi^2}{[\mathbb{E}_* \psi']^2}\right).$$

The AIF under these assumptions is $x \mapsto [\mathbb{E}_* \psi']^{-1} \psi(x)$.

Thus, the optimal location estimator ψ is the (unique?) solution to the optimisation problem:

$$\text{Minimise: } \frac{\mathbb{E}_* \psi^2}{[\mathbb{E}_* \psi']^2}, \text{ subject to: } \sup_x \left| \frac{\psi(x)}{\mathbb{E}_* \psi'} \right| \leq c,$$

for some threshold c. Immediately we see that this optimisation problem is homogeneous in ψ: ψ solves the problem if and only if $\alpha\psi$ solves it, for any constant α. We therefore add a further constraint

$$\mathbb{E}_* \psi' = 1. \tag{22}$$

We also need (see (21)) the constraint

$$\mathbb{E}_* \psi = 0. \tag{23}$$

The Lagrangian is

$$\mathcal{L}(\psi, \lambda, \mu) = \mathbb{E}_* \psi^2 + \lambda \mathbb{E}_* \psi + \mu(\mathbb{E}\psi' - 1),$$

subject to $\|\psi\|_\infty = \sup_x |\psi(x)| \leq c$. With some work, we can solve this and get that the optimal *psi*:

$$\psi(x) = \left[-\frac{1}{2}\lambda - \frac{1}{2}\mu \frac{p_*'(x)}{p_*(x)} \right]_c^c,$$

where λ and μ must be solved by constraints (22) and (23) and the notation

$$[y]_d^c = \begin{cases} y & \text{if } d \leq y \leq c, \\ c & \text{if } y > c, \\ d & \text{if } y < d. \end{cases}$$

If p_* is the standard Normal, then $\frac{p_*'(x)}{p_*(x)} = -x$ and by symmetry we would require $\lambda = 0$. Then the optimal robust estimator reduces to the Huber estimator (i.e. trimmed mean, see Example 1.4):

$$\psi(x) = [x]_c^c.$$

More details are in week 5's Section.

8 Lecture 18/2

8.1 One step estimators

add-on The reference for this lecture is section 5.7 of [vdV].

Limitations of M- and Z-estimators:

1. A maximum of the criterion function may not exist or it may not be unique. Similarly, a root of the estimating may not exist or be unique. (Note we need uniqueness of the theoretical root/maximum for consistency.)

2. Even if it does uniquely exist, the maximum/root may be difficult to compute.

The one-step method sidesteps these problems. Despite its name, it is a two-stage procedure:

1. Determine a preliminary estimator $\tilde{\theta}_n$ which is 'reasonably good' (i.e. within $n^{-1/2}$ distance from the true θ^* in some sense which we will make precise).

2. Apply some carefully crafted function f to obtain an estimator
$$\hat{\theta}_n = f(\tilde{\theta}_n),$$
which has the same variance as the corresponding M- and Z-estimation problem while avoiding the limitations above.

As in Z-estimation, our goal is to find a root $\psi_n(\theta) = 0$. Let $\tilde{\theta}_n$ be a preliminary estimator. Informally, the *one step estimator* $\hat{\theta}_n$ is the solution to

$$\psi_n(\tilde{\theta}_n) + \dot{\psi}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) = 0,$$

where $\dot{\psi}_n(\tilde{\theta}_n)$ is the derivative of $\psi_n(\theta)$ with respect to θ, evaluated at $\theta = \theta_n$. So $\hat{\theta}_n$ can be thought of as a single Newton-Raphson (NR) update, starting from $\tilde{\theta}_n$ (hence the name 'one-step').

If it is possible to invert $\dot{\psi}_n(\tilde{\theta}_n)$ then the solution is

$$\hat{\theta}_n = \tilde{\theta}_n - \left[\dot{\psi}_n(\tilde{\theta}_n)\right]^{-1} \psi_n(\tilde{\theta}_n).$$

An intuitive idea is that to numerically solve $\psi_n(\theta) = 0$ starting with a reasonable approximate solution $\tilde{\theta}_n$, recursively applying NR multiple times would mean the resulting estimator is close to the root and hence a better estimator. Yet the one-step estimator does a single NR update. It turns out that there is no asymptotic improvement from applying multiple NR updates. This shows that for large sample sizes, multiple NR updates are not very useful (although for small sample sizes, they may be).

8.1.1 Set-up

Consider a parametric (aka well-specified) setting where $X_1, \ldots, X_n \overset{iid}{\sim} p_{\theta_0}$. Suppose the estimating equations $\psi_n(\theta) = 0$ are given. Assume that there exists a non-singular matrix $\dot{\psi}_0$ such that

$$\sup_{\|\theta-\theta_0\|<M/\sqrt{n}} \left\| \sqrt{n}\left(\psi_n(\theta) - \psi_n(\theta_0)\right) - \dot{\psi}_0 \sqrt{n}(\theta - \theta_0) \right\| \xrightarrow{\mathbb{P}_{\theta_0}} 0, \tag{24}$$

for every constant M.

The LHS of (24) is a generalisation of the first-order Taylor expansion where $\dot{\psi}_0$ takes the place of the derivative.

Definition 8.1. Given a preliminary estimator $\tilde{\theta}_n$ and a sequence $\dot{\psi}_{n,0}$ of non-singular random matrices which are consistent for some matrix $\dot{\psi}_0$,

$$\hat{\theta}_n = \tilde{\theta}_n - \dot{\psi}_{n,0}^{-1}\psi_n(\tilde{\theta}_n),$$

is the *one step estimator*.

Definition 8.2. A set of random vectors $\{X_\alpha : \alpha \in A\}$ is *uniformly tight* (or equivalently, *bounded in probability*, or $O_{\mathbb{P}}(1)$) if for all $\epsilon > 0$, there exists M such that

$$\sup_{\alpha \in A} \mathbb{P}\left(\|X_\alpha\| > M\right) < \epsilon.$$

If $\tilde{\theta}_n$ is an estimator for θ_0, then $\tilde{\theta}_n$ is $\sqrt{n}$*-consistent* if $\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)$ is uniformly tight.

add-on See Assignment 2, Question 5, Part (a) for some results about bounded in probability.

$\sqrt{n}$-consistency formalises the intuitive notion that the preliminary estimator $\tilde{\theta}_n$ must be within $\sqrt{n}$ distance from the true θ_0.

8.1.2 Asymptotic theorem for one step estimators

Theorem 8.3. *Suppose:*

1. $\sqrt{n}\psi_n(\theta_0)$ *converge in distribution to some valid random variable;*

2. *there exists* $\dot{\psi}_0$ *which satisfies assumption (24);*

3. *The one step estimator* $\hat{\theta}_n$ *is such that*

   (a) *its preliminary estimator* $\tilde{\theta}_n$ *is* $\sqrt{n}$*-consistent;*

   (b) $\dot{\psi}_{n,0} \xrightarrow{\mathbb{P}} \dot{\psi}_0$.

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = -\dot{\psi}_0^{-1}\sqrt{n}\psi_n(\theta_0) + o_{\mathbb{P}}(1).$$

The conclusion of this theorem is very similar to that of the Z-estimator (Theorem 7.1). The advantage is that we don't need to know the variance of $\tilde{\theta}_n$, yet we can obtain the asymptotic variance of $\hat{\theta}_n$. Moreover, the asymptotic variance of $\hat{\theta}_n$ is the same as the Z-estimator using estimating equations $\psi_n(\theta) = 0$. Hence, the one step estimator is useful in scenarios where the Z-estimator is not well defined or we can't find a root easily.

"The Theorem shows that $\hat{\theta}_n$ is $\sqrt{n}$-consistent so you could use a one step estimator as the preliminary estimator $\tilde{\theta}_n$ (i.e. do multiple NR updates), but this Theorem shows you will not improve the asymptotic variance by doing this." The single update is enough for large sample sizes (for small sample sizes, it may be worth doing multiple steps).

Proof. Our goal is to study $\sqrt{n}\dot{\psi}_0\left(\hat{\theta}_n - \theta_0\right)$. We will start by looking at

$$\begin{aligned}
\sqrt{n}\dot{\psi}_{n,0}\left(\hat{\theta}_n - \theta_0\right) &= \sqrt{n}\dot{\psi}_{n,0}\left(\tilde{\theta}_n - \dot{\psi}_{n,0}^{-1}\psi_n(\tilde{\theta}_n) - \theta_0\right) \\
&= \sqrt{n}\dot{\psi}_{n,0}\left(\tilde{\theta}_n - \dot{\psi}_{n,0}^{-1}\psi_n(\tilde{\theta}_n) + \dot{\psi}_{n,0}^{-1}\psi_n(\theta_0) - \dot{\psi}_{n,0}^{-1}\psi_n(\theta_0) - \theta_0\right) \\
&= \dot{\psi}_{n,0}\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) - \sqrt{n}\left(\psi_n(\tilde{\theta}_n) - \psi_n(\theta_0)\right) - \sqrt{n}\psi_n(\theta_0) \\
&= (\dot{\psi}_{n,0} - \dot{\psi}_0)\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) - \sqrt{n}\psi_n(\theta_0) + o_{\mathbb{P}}(1) \\
&= -\sqrt{n}\psi_n(\theta_0) + o_{\mathbb{P}}(1) \qquad (25)
\end{aligned}$$

where the second line is valid since we assume $\psi_n(\theta_0)$ is not infinite a.s., even as $n \to \infty$; the fourth line follows since by assumption 2.:

$$\sqrt{n}\left(\psi_n(\tilde{\theta}_n) - \psi_n(\theta_0)\right) - \dot{\psi}_0\sqrt{n}(\tilde{\theta}_n - \theta_0) = o_{\mathbb{P}}(1),$$

and the first term of the fourth line goes to zero in probability by Slutsky's theorem since $\dot{\psi}_{n,0} \xrightarrow{\mathbb{P}} \dot{\psi}_0$ and $\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)$ is uniformly bounded.

(25) shows that $\hat{\theta}_n$ is also $\sqrt{n}$-consistent for θ_0, so

$$\sqrt{n}\dot{\psi}_0\left(\hat{\theta}_n - \theta_0\right) - \sqrt{n}\dot{\psi}_{n,0}\left(\hat{\theta}_n - \theta_0\right) = o_{\mathbb{P}}(1),$$

by similar reasoning to above. Combine this with (25) to get the required result. □

8.1.3 Examples of the one step estimator's utility

Example 8.4. Two concrete examples where one step estimators are a better choice than maximising likelihood:

1. Suppose $X_1, \dots, X_n \overset{iid}{\sim} \text{Cauchy}(\theta, \pi)$. ($\theta$ is the location parameter and π the scale.) Then
$$p_{\theta,\pi}(x) = \frac{1}{\pi(1 + (x - \theta)^2)}.$$
The profile log-likelihood (of θ, with π fixed) often has multiple roots. In fact, the number of roots behaves asymptotically like $2\text{Pois}(1/\pi) + 1$. A good choice for the preliminary estimator $\tilde{\theta}_n$ is the median. (See Assignment 2, Question 5; example 5.50 of [vdV].)

2. Mixture of densities (likelihood diverges): Let f and g be given probability densities with common support $\mathbb{R}$. Given the parameter $\boldsymbol{\theta} = (\mu, \nu, \sigma, \tau, p)$, draw $X_1, \dots, X_n$ iid from the distribution with density
$$x \mapsto pf\left(\frac{x - \mu}{\sigma}\right)\frac{1}{\sigma} + (1 - p)g\left(\frac{x - \nu}{\tau}\right)\frac{1}{\tau}.$$
The likelihood is unbounded on the unrestricted parameter space, so the MLE is undefined. You could try an M- or Z-estimator (but what to use as the criterion function/estimation equation?). Or try a one step estimation with a MoM estimator as the preliminary estimator.

Typically, the preliminary estimator $\tilde{\theta}_n$ is chosen to be a MoM estimator or a robust estimator (e.g. the median).

8.2 Roadmap: Asymptotic Optimality and Risk

How do we choose among all the estimators we've developed? We need a notion of asymptotic optimality. In Stat210, we looked at finite-sample optimality (e.g. admissibility, minimaxity, Bayes risk), but this was a case-by-case study. Asymptotics typically allows for unifying many smooth parametric models – this is the case for optimality as well.

The roadmap for the next few lectures is to start by trying to extend finite sample admissibility to an asymptotic notion; by way of the example of the Hodges estimator, we will show the natural extension breaks; and this will lead us to a proper definition of asymptotic optimality.

9 Lecture 23/2

9.1 Asymptotic Admissibility?

Two fundamental questions: Can we design 'optimal' estimators? Can we design 'asymptotically optimal' estimators?

For this discussion of optimality, we will restrict to the parametric setting. Given $X_1, \dots, X_n \overset{iid}{\sim} p_\theta$ with $\theta \in \mathcal{H}$, our goal is to estimate $g(\theta)$.

Definition 9.1. Given a risk function $R(\delta_n, g(\theta)) = \mathbb{E}_\theta \left(l(\delta_n(X), g(\theta)) \right)$, an estimator δ_n is *inadmissible* (in a class $\mathcal{C}$ of estimators) if δ_n is *dominated* by another estimator – i.e. there exists another estimator $\tilde{\delta}_n \in \mathcal{C}$ with

$$R(\tilde{\delta}_n, g(\theta)) \leq R(\delta_n, g(\theta)),$$

for all θ, and strict inequality for at least one θ.

How can we extend this notion to the asymptotic setting? A natural definition is to say δ_n is asymptotically inadmissible if there exists $\tilde{\delta}_n$ with

$$\lim_{n\to\infty} \frac{R(\tilde{\delta}_n, g(\theta))}{R(\delta_n, g(\theta))} \leq 1,$$

for all θ and strict inequality for at least one θ. (add-on Technically we should probably use lim sup instead of lim.) This is just a notion to motivate our discussion – it's not a definition found in the literature since we will see it is not useful in the following example.

Example 9.2. Given $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$, consider the Hodges estimator

$$\tilde{\delta}_n = \begin{cases} \bar{X}_n & \text{if } \left|\bar{X}_n\right| \geq n^{-1/4}, \\ 0 & \text{otherwise.} \end{cases}$$

Also, define $\delta_n = \bar{X}_n$ to be the sample mean. Consider the loss $l(u, v) = (u - v)^2$. Then

$$\lim_{n\to\infty} nR(\delta_n, \theta) = \lim_{n\to\infty} n\mathrm{Var}_\theta(\bar{X}_n) = 1,$$

for all θ.

What is the risk of the Hodges estimator $\tilde{\delta}_n$? Intuition: when $\theta = 0$, $\bar{X}_n$ converges to zero at the rate $\sqrt{n}$ (by the CLT) – so it converges faster than $n^{-1/4}$. This means $\bar{X}$ will be in the interval $[-n^{-1/4}, n^{1/4}]$ with probability going to one. Hence $nR(\tilde{\delta}_n, 0) \to 0$. Outside a local neighbourhood of $\theta = 0$, $\tilde{\delta}_n$ behaves like the sample mean δ_n. We can make this argument rigorous (see Assignment 2, Question 6) and show

$$\lim_{n\to\infty} nR(\tilde{\delta}_n, \theta) = \begin{cases} 1 & \text{if } \theta \neq 0, \\ 0 & \text{if } \theta = 0. \end{cases}$$

Hence the sample mean δ_n is dominated by the Hodges estimator $\tilde{\delta}_n$.

What about the finite sample picture? For the sample mean $R(\delta_n, \theta) = 1$ everywhere, for all n.

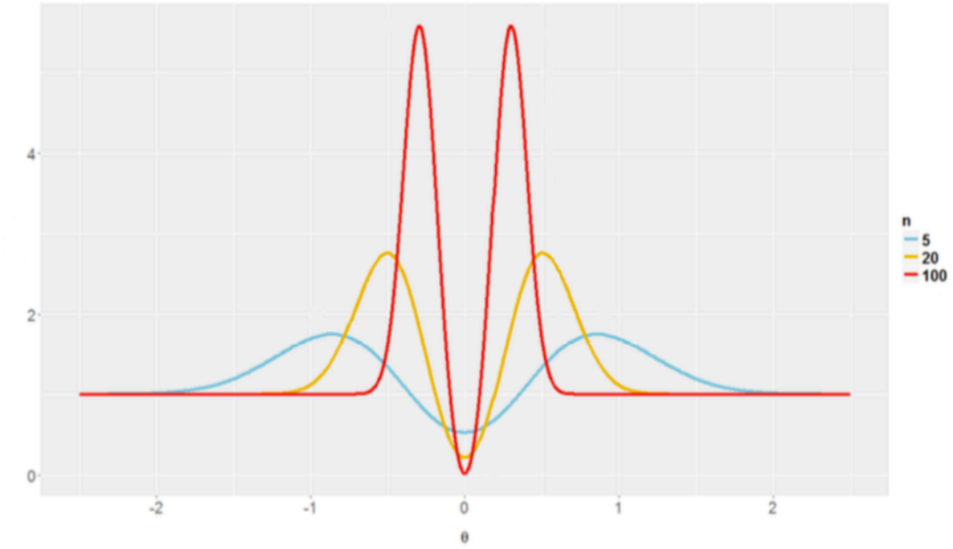


Figure 1: Risk function of the Hodges estimator for increasing values of n.

For the Hodges estimator,

$$\sup_{\theta \in \mathbb{R}} R(\tilde{\delta}_n, \theta) \to \infty, \tag{26}$$

as $n \to \infty$ (see Assignment 2, Question 6 for calculations). The spike at $n^{-1/4}$ goes to infinity as $n \to \infty$. So $\tilde{\delta}_n$ has really bad minimax risk. Our definition of asymptotic admissibility is not a unifying notion since it doesn't capture/explain this phenomenon. (This is also an argument for looking at minimaxity and not just admissibility.)

A natural fix is to discard sets of measure zero from consideration.

Definition 9.3. $\tilde{\delta}_n$ is *asymptotically admissible* if, for all δ_n satisfying

$$\lim_{n\to\infty} \frac{R(\delta_n, g(\theta))}{R(\tilde{\delta}_n, g(\theta))} \leq 1,$$

for all θ, strict inequality only occurs for θ in a set of Lebesgue measure zero.

(This definition is not in the literature – instead there is an analogous notion, which we might cover later in class. add-on This definition doesn't really make sense since it assumes that the limit always exists.)

Under this definition, $\delta_n = \bar{X}_n$ from Example 9.2 is asymptotically admissible.

Our primary concern with the old definition of asymptotic admissibility is that it didn't capture the finite sample phenomenon of the Hodges estimator. This is a crucial point since a basic requirement of asymptotics is that they reflect finite sample behaviour. Yet this new definition also doesn't address this concern. The fix is local asymptotic analysis.

9.2 Local asymptotic analysis

Why was the previous asymptotic analysis lacking? It missed important behaviour which occurred locally around zero. From this emerges a natural thought: can we study the normalised risk locally around θ_0?

Definition 9.4. Given parameter space $\mathcal{H} \subset \mathbb{R}^p$, fix some $\alpha > 0$ and $h \in \mathbb{R}^p$. The *local risk* at θ_0 of an estimator δ_n (in the direction h with scale α), is given by

$$\lim_{n\to\infty} nR(\delta_n, g(\theta_0 + h/n^\alpha)) = \lim_{n\to\infty} n\mathbb{E}_{\theta_0+h/n^\alpha}\left[l(\delta_n, g(\theta_0 + h/n^\alpha))\right].$$

The choice of α matters – some α will capture finite sample phenomenon and some α will not. We will take it as given that $\alpha = 1/2$ captures finite sample behaviour for most smooth parametric models. (See Section.)

Example 9.5 (continuation of Example 9.2). Consider $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(h/\sqrt{n}, 1)$. With some work, we can compute

$$[\lim_{n\to\infty} nR(\tilde{\delta}_n, h/n^\alpha) = h^2,$$

while the MLE $\bar{X}_n$ has local risk 1. So for $|h| < 1$, the Hodges estimator is 'better'; otherwise the MLE is.

We used $n^{-1/4}$ in the Hodges estimator. What happens if we change the exponent? How will that change the local risk? Define

$$\delta'_n = \begin{cases} \bar{X}_n & \text{if } \left|\bar{X}_n\right| \geq n^{-2.5}, \\ 0 & \text{otherwise.} \end{cases}$$

For $\alpha = 1/2$, the local risk captures the behaviour of δ'_n (i.e. it gives similar answers to $\tilde{\delta}_n$). For other α¡ the local risk goes to infinity or doesn't capture the local behaviour of δ'_n¿ This is meant to be an example to illustrate that $\alpha = 1/2$ is the right value to use.

There are two takeaways from this discussion: There is a need for local asymptotic analysis. And the scale α of the local analysis is crucial. We should take as given that $\alpha = 1/2$ is the right scale as long as you are in smooth parametric families (e.g. ones that satisfy the assumptions of the asymptotic normality of MLE theorem.)

9.3 Local asymptotic distribution

We have consider local analysis in terms of risk. But we may be interested in other local properties. Specifically, can we understand the asymptotic distribution of

$$\sqrt{n}\left(\delta_n - g(\theta_0 + h/\sqrt{n})\right), \tag{27}$$

where $X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0+h/\sqrt{n}}$? Understanding the asymptotic distribution in local neighbourhoods of θ_0 will allow for a more refined analysis/capture more information, as compared to a local asymptotic risk analysis.

Can we study the local asymptotic distribution with tools we've already seen? No, because the data generating distribution is changing with n. (Aside: this is probably the first time that we are really touching on a link to high dimensional research, where $\dim(\theta)$ can grow – and hence the model choice changes – as a function of n.) But we can transform (27) into something that looks more familiar:

$$\sqrt{n}\left(\delta_n - g(\theta_0 + h/\sqrt{n})\right) = \sqrt{n}\left(\delta_n - g(\theta_0)\right) + \sqrt{n}\left(g(\theta_0) - g(\theta_0 + h/\sqrt{n})\right).$$

We can understand the second term:

$$\sqrt{n}\left(g(\theta_0) - g(\theta_0 + h/\sqrt{n})\right) = \sqrt{n}\frac{g(\theta_0) - g(\theta_0 + h/\sqrt{n})}{h/\sqrt{n}}\frac{h}{\sqrt{n}} \xrightarrow{n\to\infty} -A_{\theta_0}h,$$

where A_{θ_0} is the Jacobian of $g(\theta)$ at θ_0. We would know the first term if the true data generating parameter was θ_0. But to understand it under $p_{\theta_0+h/\sqrt{n}}$, we will need to develop some theory of 'tilted measures'.

10 Lecture 25/2

10.1 Preview

Recall the set-up of section 9.3: Given $X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0+h/\sqrt{n}}$ and an estimator δ_n of $g(\theta_0)$, we want to understand the asymptotic distribution of $\sqrt{n}\left(\delta_n - g(\theta_0 + h/\sqrt{n})\right)$. In the previous lecture, we reduced this to understanding the asymptotic distribution of $\sqrt{n}\left(\delta_n - g(\theta_0)\right)$ under $p_{\theta_0+h/\sqrt{n}}$.

Let $T_n = \sqrt{n}\left(\delta_n - g(\theta_0)\right)$. Under some assumptions we know that $T_n \xrightarrow[p_{\theta_0}]{d} T$. By the Portmanteau theorem, this is equivalent to

$$\int f\left(T_n\right) p_{\theta_0}^{\otimes n} \to \int f(t) dP(t),$$

for all bounded continuous functions f, where $P(t)$ is the law of T and $p_{\theta_0}^{\otimes n}$ is the n-fold product measure with each component p_{θ_0}. We want to study the asymptotic distribution of T_n. By the Portmanteau theorem, this is equivalent to studying the convergence of $\int f(T_n) dp^{\otimes n}_{\theta_0+h/\sqrt{n}}$. If $p^{\otimes n}_{\theta_0+h/\sqrt{n}}$ is absolutely continuous with regard to $p^{\otimes n}_{\theta_0}$ then

$$\int f(T_n) dp^{\otimes n}_{\theta_0+h/\sqrt{n}} = \int f(T_n) \frac{dp^{\otimes n}_{\theta_0+h/\sqrt{n}}}{dp^{\otimes n}_{\theta_0}} dp^{\otimes n}_{\theta_0}.$$

If we know that

$$\left(T_n, \frac{dp^{\otimes n}_{\theta_0+h/\sqrt{n}}}{dp^{\otimes n}_{\theta_0}} \right) \xrightarrow[p_{\theta_0}]{d} (T, V),$$

then we would expect that

$$\int f(T_n) \frac{dp^{\otimes n}_{\theta_0+h/\sqrt{n}}}{dp^{\otimes n}_{\theta_0}} dp^{\otimes n}_{\theta_0} \to \int f(t) v \; dP(T, V), \tag{28}$$

where $P(T, V)$ is the law of (T, V). So we would have achieved our goal of understanding the asymptotic distribution of T_n under $p_{\theta_0+h/\sqrt{n}}$:

$$T_n \xrightarrow[p^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}}]{d} T',$$

where the law of T' is given by $P_{T'}(B) = \mathbb{E}_P \left[\mathbb{1}\{T \in B\} \, V \right]$.

For this to work, we assumed absolutely continuity of $p^{\otimes n}_{\theta_0+h/\sqrt{n}}$ with respect to $p^{\otimes n}_{\theta_0}$. But even if we don't have absolutely continuity for some n, it will still work as long as we have absolutely continuity for large n – i.e. as long as we have some notion of "asymptotic absolutely continuity". This notion is called contiguity.

10.2 Some necessary measure theory

To rigorously prove (28), we need to develop some measure theory. Throughout this section, we will use P and Q to denote two measures – as a concrete example, take Q to be some measure on $\mathbb{R}^k$ and P to be the Lebesgue measure on the same space.

Definition 10.1. Let P and Q be two measures defined on a measure space $(\Omega, \mathcal{A})$. Q has a *density* f with respect to P if

$$Q(A) = \int_A f dP,$$

for all measurable sets $A \in \mathcal{A}$. In this context P is called the *base measure.*

When does a density f exist? The critical requirement is absolute continuity.

Definition 10.2. Let P and Q be two measures on the measure space $(\Omega, \mathcal{A})$. We say Q is *absolutely continuous* with respect to P and write $Q \ll P$ if

$$P(A) = 0 \Rightarrow Q(A) = 0,$$

for all measurable sets A.

If P and Q have densities f_P and f_Q, absolute continuity requires the support of f_Q to lie inside the support of f_P. Absolutely continuity of measures is related to the notion of absolutely continuous functions: If μ is a finite measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ then μ is absolutely continuous with respect to the Lebesgue measure if and only if μ's CDF is absolutely continuous. This shows that a probability measure F has a density if and only if F is absolutely continuous with respect to to the Lebesgue measure[‡].

10.2.1 The Radon-Nikodym derivative

Theorem 10.3 (Radon-Nikodym). *Let P and Q be two σ-finite[§] measures on $(\Omega, \mathcal{A})$ and $Q \ll P$. Then there exists a measurable function $f : \Omega \to [0, \infty]$ (i.e. a density) such that for any measurable set $A \in \mathcal{A}$,*

$$Q(A) = \int_A f dP.$$

f is called the *Radon-Nikodym derivative.* When P and Q have densities (with respect to same base measure – typically Lebesgue measure), f is simply the ratio of the densities (which exists by absolute continuity). Hence a common name for f in statistics is the *likelihood ratio.*

If f and $\tilde{f}$ are two Radon-Nikodym derivatives, then $f = \tilde{f}$ a.e. with respect to P.

[‡] Continuity of F is not sufficient for a density to exist. The canonical counterexample is the Cantor function. Differentiability of F is a sufficient but not necessary condition.

[§] A measure P on $(\Omega, \mathcal{A})$ is σ-finite if Ω can be covered by countably many measurable sets with finite, or equivalently if there exists a strictly positive measurable function $f : \Omega \to \mathbb{R}$ whose integral with respect to P is finite.

How do we construct the Radon-Nikodym derivative? If Ω is a metric space (maybe with some regularity conditions) and Q is a finite measure then

$$\frac{dQ}{dP}(\omega) := \lim_{r \to 0} \frac{Q(B(\omega, r))}{P(B(\omega, r))}, \tag{29}$$

exists P-a.e., where $B(\omega, r)$ is the open ball centred at ω with radius r. (This is Theorem 35.7 of [Bil12].) Further, if $Q \ll P$ then $\frac{dQ}{dP}$ equals the Radon-Nikodym derivative P-a.e.

Claim 10.4. *For any measurable function h and any measurable set A,*

$$\int_A h dQ = \int_A h \frac{dQ}{dP} dP,$$

assuming $Q \ll P$, (so that $\frac{dQ}{dP}$ equals the Radon-Nikodym derivative.)

The proof is a simple application of InSiPoD.

10.2.2 Lebesgue's decomposition theorem

(Jordan's decomposition:) A probability measure Q on $\mathbb{R}$ can always be decomposed into an absolutely continuous part and a discrete[¶] part:

$$\begin{aligned} Q(A) &= Q_c(A) + Q_d(A) \\ &= \int_A f d\mu(x) + Q_d(A), \end{aligned}$$

where f is the density of Q_c and μ is the Lebesgue measure. This result can be generalised beyond the Lebesgue measure.

Definition 10.5. Let P and Q be two measures on $(\Omega, \mathcal{A})$. P and Q are *singular* with respect to to each other (denoted $P \perp Q$) if Ω can be partitioned into two disjoint sets A and B such that $P(B) = Q(A) = 0$.

[¶] A random variable is discrete if its support is countable.

Theorem 10.6 (Lebegue's decomposition)**.** *Suppose P and Q are two σ-finite measures on $(\Omega, \mathcal{A})$. Then we can decompose*

$$Q = Q_a + Q_s,$$

where $Q_a \ll P$ and $Q_s \perp P$. Further, this decomposition is unique.

How can we characterise this decomposition? (Lemma 6.2 from [vdV]:) If P and Q are absolutely continuous with respect to a base measure μ, with densities p and q, then

$$\begin{aligned} Q_a(A) &= Q(A \cap \{p(x) > 0\}) \\ Q_s(A) &= Q(A \cap \{p(x) = 0\}). \end{aligned}$$

For any σ-finite measures P and Q,

$$\begin{aligned} Q(A) &= Q_a(A) + Q_s(A) \\ &= \int_A f dP + Q_s(A), \end{aligned}$$

where f is the Radon-Nikodym derivative of Q_a with respect to P.

Claim 10.7. *On the set (of P-measure 1) where $\frac{dQ}{dP}$ (29) exists, it is equal to the Radon-Nikodym derivative f of Q_a with respect to P.*

Hence, from herein *we will use $\frac{dQ}{dP}$ to denote the Radon-Nikodym derivative f* of Q_a with respect to P and we will call $\frac{dQ}{dP}$ the likelihood ratio.

We can generalise Claim 10.4 when Q is not necessarily absolutely continuous with respect to P:

Claim 10.8.

$$\int_A h \frac{dQ}{dP} dP \leq \int_A h dQ,$$

for any σ-finite measures Q and P defined on the same measure space and any measureable set A.

10.2.3 Contiguity

Contiguity is the generalisation of absolute continuity to asymptotics.

Definition 10.9. Let $(\Omega_n, \mathcal{A}_n)$ be a sequence of measure spaces, each equipped with probability measures P_n and Q_n. Then $\{Q_n\}$ is *contiguous* with respect to $\{P_n\}$ (denoted $Q_n \lhd P_n$) if

$$P_n(A_n) \to 0 \Rightarrow Q_n(A_n) \to 0,$$

for every sequence $\{A_n\}$ of measurable sets.

Example 10.10. Fix $P_n = \text{Unif}[0,1]$.

1. $Q_n = \text{Unif}[n, n+1]$ is not contiguous with respect to P_n since $P_n(A_n) = 0$ and $Q_n(A_n) = 1$ for $A_n = [n, n+1]$.

2. $Q_n = \text{Unif}[0.5 + 1/n, 1.5 + 1/n]$ is not contiguous with respect to P_n. Why? Take $A_n = [1, 1.5]$. Then $P_n(A_n) = 0$ and $Q_n(A_n) = 0.5$ for $n \geq 2$.

3. $Q_n = \text{Unif}[1/n, 1 + 1/n]$ is contiguous with respect to P_n. To prove this, fix A_n and partition $\mathbb{R}$ according to the supports' of P_n and Q_n: Define

$$\begin{aligned} B_n &= A_n \cap [1/n, 1] \\ C_n &= A_n \cap [0, 1/n] \\ D_n &= A_n \cap [1, 1/n] \\ E_n &= A_n \setminus (B_n \cup C_n \cup D_n) . \end{aligned}$$

   To show $Q_n(A_n) \to 0$, show that the Q_n-measure of B_n through E_n goes to zero.

4. $Q_n = \text{Unif}[0.5 + 1/n, 1 + 1/n]$ is contiguous with respect to P_n. To prove this, use the same idea as 3.

5. Now define $P_n = \mathcal{N}(0,1)$ and $Q_n = \mathcal{N}(n,1)$. Then $Q_n \not\lhd P_n$ – take $A_n = [n - c, n + c]$ for some constant c as the counterexample.

Loosely: a necessary condition for contiguity is that the support of Q_n converges to a subset of the support of P_n. Further, if P_n and Q_n have the same support, their centres of mass cannot move away from each other.

Example 5. shows that it is possible that $Q_n \ll P_n$ for all n but $Q_n \ntriangleleft P_n$. Example 4. shows that it is possible for $Q_n \not\ll P_n$ for all n but $Q_n \triangleleft P_n$.

11 Lecture 2/3

11.1 Local asymptotic analysis

Theorem 11.1 ("Le Cam's main theorem")**.** *Let P_n and Q_n be two sequences of probability measures on $(\Omega_n, \mathcal{A}_n)$ and X_n be a sequence of random vectors (of constant dimension k). Suppose*

(i) $Q_n \triangleleft P_n$*;*

(ii) $\left(X_n, \frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} (X, V)$[‖]*.*

Then

$$X_n \xrightarrow[Q_n]{d} L,$$

that is, the law of X_n induced by Q_n converges to the law

$$Q_L(B) = \mathbb{E}_P\left[\mathbb{1}\{X \in B\}\, V\right] = \int \mathbb{1}\{x \in B\}\, v dP(x, v), \tag{30}$$

(where the expectation is under the joint distribution P of (X, V) from (ii)).

Notation: $X_n \xrightarrow[P_n]{d} X$ means that the law of X_n induced by the measure P_n converges weakly to the law of X.

Analogy: If Q and P are probability measures with $Q \ll P$, then

$$\int f dQ = \int f \frac{dQ}{dP} dP,$$

[‖] This may seem confusing since on the face of it, $\frac{dQ_n}{dP_n}$ doesn't appear to be a random – so how can it converge to a random variable? But $\frac{dQ_n}{dP_n}$ is a random variable – it is a function $\Omega \to \mathbb{R}$!

for all measurable f. Choose $f(x) = \mathbb{1}\{x \in B\}$. Then $Q(X \in B) = \mathbb{E}_P\left[\mathbb{1}\{X \in B\}\frac{dQ}{dP}\right]$, which is very similar to (30). More generally,

$$\int f dQ = \int f \frac{dQ_a}{dP} dP + \int f dQ_s,$$

where $Q = Q_a + Q_s$ is the Lebesgue decomposition. Then

$$Q(X \in B) = \mathbb{E}_P\left[\mathbb{1}\{X \in B\}\frac{dQ_a}{dP}\right] + Q_s(X \in B).$$

The intuition is that on the support of $Q_{n,s}$, the measure P_n is zero – that is, $Q_{n,s}(B) = Q_n(B \cap p_n(x) = 0)$ (where p_n is the density of P_n) yet $P_n(B \cap p_n(x) = 0) = 0$. So by contiguity $Q_{n,s}(B)$ must shrink to zero, for all B. This then implies

$$\begin{aligned}\lim_{n\to\infty} Q_n(X_n \in B) &= \lim_{n\to\infty} \mathbb{E}_{P_n}\left[\mathbb{1}\{X_n \in B\}\frac{dQ_{n,a}}{dP_n}\right] + \lim_{n\to\infty} Q_{n,s}(X_n \in B) \\ &= \lim_{n\to\infty} \mathbb{E}_{P_n}\left[\mathbb{1}\{X_n \in B\}\frac{dQ_{a,n}}{dP_n}\right].\end{aligned}$$

We want to use this theorem with $P_n = p_{\theta_0}^{\otimes n}$ and $Q_n = p_{\theta_0 + \frac{h}{\sqrt{n}}}^{\otimes n}$ where $X_n = \sqrt{n}\left(\hat{\theta}_n - \theta\right)$ or $X_n = \sqrt{n}\left(\hat{\theta}_n - \theta - \frac{h}{\sqrt{n}}\right)$. We would then need to show that $p_{\theta_0+\frac{h}{\sqrt{n}}}^{\otimes n} \lhd p_{\theta_0}^{\otimes n}$ and that the limiting distribution of $\left(X_n, \frac{dp_{\theta_0+h/\sqrt{n}}^{\otimes n}}{dp_{\theta_0}^{\otimes n}}\right)$ exists.

Proof. First we need to prove that Q_L is a valid probability law (that is $\int dQ_L = 1$). (This is where contiguity is used.) We omit the proof of this.

Second, use the portmanteau theorem (Lemma 2.2 of [vdV]), which gives an equivlanet characterisation of convergence in distribution: $X_n \xrightarrow[Q_n]{d} L$ if and only if

$$\liminf_{n\to\infty} \mathbb{E}_{Q_n}\left[f(X_n)\right] \geq \mathbb{E}_{Q_L}\left[f(X)\right],$$

for every non-negative continuous measurable function f.

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}_{Q_n}\left[f(X_n)\right] &\geq \liminf_{n\to\infty} \mathbb{E}_{P_n}\left[f(X_n)\frac{dQ_n}{dP_n}\right] \\ &\geq \mathbb{E}_P\left[f(X)V\right] \\ &= \mathbb{E}_{Q_L}\left[f(L)\right]\end{aligned}$$

where the first line uses Claim 10.8; the second line uses the portmanteau theorem and the convergence of $\left(X_n, \frac{dQ_n}{dP_n}\right)$; and the final line follows by InSiPoD. □

11.2 Equivalent characterisations of contiguity (Le Cam's first lemma)

How can we show contiguity in practice? Typically we will use the following lemma.

Lemma 11.2 (Le Cam's first lemma (Lemma 6.4 of [vdV]))**.** *Suppose P_n and Q_n are sequences of probability measures on measure spaces $(\Omega_n, \mathcal{A}_n)$. Then the following statements are equivalent:*

1) $Q_n \lhd P_n$*;*

2) $\frac{dQ_n}{dP_n} \underset{P_n}{\overset{d}{\rightsquigarrow}} V$ *along a subsequence then* $\mathbb{E}[V] = 1$*;*

3) If $\frac{dP_n}{dQ_n} \underset{Q_n}{\overset{d}{\rightsquigarrow}} U$ *along a subsequence then* $P(U > 0) = 1$*;*

4) For any statistics $T_n : \Omega_n \to \mathbb{R}^k$*, if* $T_n \xrightarrow{P_n} 0$*, then* $T_n \xrightarrow{Q_n} 0$*.*

The proof is given in Section 6.

Corollary 11.3 (an application of Le Cam's first lemma)**.** *If P_n and Q_n are such that*

$$\log \frac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} \mathcal{N}(\mu, \sigma^2),$$

then

(i) $Q_n \lhd P_n$*; and*

(ii) $P_n \lhd Q_n$ *if and only if* $\mu = -\frac{1}{2}\sigma^2$*.*

(If $Q_n \lhd P_n$ and $P_n \lhd Q_n$ then we say P_n and Q_n are *mutually contiguous* and we write $P_n \lhd\rhd Q_n$.)

Proof. Use CMT to get that

$$\frac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} \exp\left[\mathcal{N}(\mu, \sigma^2)\right].$$

Then 3) from Lemma 11.2 is satisfied so $Q_n \lhd P_n$. Use 2) from Lemma 11.2 with the fact that

$$\mathbb{E}\left(\exp\left[\mathcal{N}(\mu, \sigma^2)\right]\right) = e^{\mu + \frac{1}{2}\sigma^2},$$

which equals 1 if and only if $\mu = -\frac{1}{2}\sigma^2$. □

12 Lecture 4/3

In Theorem 11.1, we were able to characterise the limiting distribution L of X_n under Q_n by $\mathbb{P}L \in B) = \mathbb{E}\left[\mathbb{1}\{X \in B\}\, V\right]$. In practice, we would often like a more concrete formulation/characterisation of L. This is the goal for this lecture. We will first study the MLE and then generalise this to a wide class of estimators.

12.1 Local asymptotic analysis of the MLE

Setup: $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$. Assume the MLE exists. We wish to study the limiting distribution of $\sqrt{n}(\hat{\theta}^{\mathrm{ML}} - \theta - h/\sqrt{n})$ under $Q_n = p_{\theta+h/\sqrt{n}}^{\otimes n}$. Since h is fixed, it suffices to study $\sqrt{n}(\hat{\theta}^{\mathrm{ML}} - \theta)$ under Q_n.

Terminology: We call $Q_n = p_{\theta+h/\sqrt{n}}^{\otimes n}$ the alternative distribution or the distribution under the local alternative. (We will see later that this local analysis has a hypothesis testing interpretation.) $P_n = p_\theta^{\otimes n}$ is called the null, or the null distribution.

Define $X_n = \sqrt{n}(\hat{\theta}^{\mathrm{ML}} - \theta)$. We want to apply Theorem 11.1 to establish the local asymptotic distribution of X_n. There are three steps to do this:

1. Check contiguity $Q_n \lhd P_n$.
2. Establish joint convergence of $\left(X_n, \frac{dQ_n}{dP_n}\right)$ under P_n.
3. Simplify the definition of L (hopefully determining it has a well-known distribution).

12.1.1 Step 1: Contiguity

We will use Corollary 11.3 (the Normal example of Le Cam's 1st Lemma). To apply the corollary, we will calculate the log likelihood ratio converges to a Gaussian under Q_n. Given $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$,

$$\begin{aligned}
\log \frac{dp_{\theta+h/\sqrt{n}}^{\otimes n}}{dp_\theta^{\otimes n}} &= \sum_{i=1}^n \left[l_i(\theta + h/\sqrt{n}) - l_i(\theta)\right] \\
&= \frac{h^\intercal}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta l_i(\theta) + \frac{h^\intercal}{2n} \sum_{i=1}^n \nabla_\theta^2 l_i(\theta) h + o_{p_{\theta_0}^{\otimes n}}(1)
\end{aligned}$$

$$\xrightarrow[p_{\theta_0}^{\otimes n}]{d} \mathcal{N}\left(-\frac{1}{2}h^\intercal I_\theta h, h^\intercal I_\theta h\right),$$

where for the second line, we assume l_n is thrice continuously differentiable so that we can Taylor expand and the third line follows since $\frac{h^\intercal}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta l_i(\theta)$ converges to $\mathcal{N}(0, h^\intercal I_\theta h)$ by the CLT and $\frac{h^\intercal}{2n}\sum_{i=1}^n \nabla_\theta^2 l_i(\theta)h$ converges to $-\frac{1}{2}h^\intercal I_\theta h$ by the WLLN.

Hence $p_{\theta+h/\sqrt{n}}^{\otimes n} \lhd\rhd p_\theta^{\otimes n}$.

We assumed a smoothness condition so that we could apply a Taylor expansion. In fact, all this is needed is QMD. (In some sense, QMD is the weakest notion that allows for this unifying parametric theory.)

Theorem 12.1. *Suppose the parameter space $\mathcal{H}$ is an open subset of $\mathbb{R}^k$ and that the model p_θ is QMD at θ_0. If $X_i \overset{iid}{\sim} p_{\theta_0}$,*

$$\left|\log\frac{dp_{\theta_0+h/\sqrt{n}}^{\otimes n}}{dp_{\theta_0}^{\otimes n}} - \frac{h^\intercal}{\sqrt{n}}\sum_{i=1}^n \eta(\theta_0, X_i) + \frac{1}{2}h^\intercal I_{\theta_0} h\right| \xrightarrow[p_{\theta_0}^{\otimes n}]{P} 0.$$

Hence $p_{\theta_0+h/\sqrt{n}}^{\otimes n} \lhd\rhd p_{\theta_0}^{\otimes n}$ by Corollary 11.3.

We omit the proof of this Theorem. Here I_{θ_0} is given as in Definition 6.5 for QMD families. Recall that under some conditions, $\eta(\theta_0, X_i) = \nabla_\theta \log p_\theta(X_i)|_{\theta=\theta_0}$ and we recover the original definition of Fisher information.

12.1.2 Step 2: Joint convergence of the MLE and the tilted measure

The goal of this step is to understand the joint asymptotic distribution of

$$\left(\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right), \log\frac{dp_{\theta+h/\sqrt{n}}^{\otimes n}}{dp_\theta^{\otimes n}}\right),$$

under $p_\theta^{\otimes n}$. In QMD families, we know the asymptotic distribution of both the MLE and the log likelihood ratio in terms of the score function $\eta(\theta, x)$ and the Fisher information I_θ. So we can determine the joint asymptotic distribution. To keep the exposition simple, we will spell this out assuming 3rd order smoothness of l_n (i.e. l_n is thrice continuously differentiable). Then we know

$$X_n = \sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right) = I_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta l_n(\theta) + o_{p_\theta^{\otimes n}}(1)$$

$$\log \frac{dp^{\otimes n}_{\theta+h/\sqrt{n}}}{dp^{\otimes n}_{\theta}} = h^\mathsf{T} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta l_n(\theta) - \frac{1}{2} h^\mathsf{T} I_\theta h + o_{p^{\otimes n}_\theta}(1).$$

Since everything on the RHS is in terms of the score statistic, dependence between X_n and $\log \frac{dp^{\otimes n}_{\theta+h/\sqrt{n}}}{dp^{\otimes n}_{\theta}}$ is taken care of. We know that $X_n \xrightarrow[p^{\otimes n}_\theta]{d} \mathcal{N}\left(0, I_\theta^{-1}\right)$ and $\log \frac{dp^{\otimes n}_{\theta+h/\sqrt{n}}}{dp^{\otimes n}_{\theta}} \xrightarrow[p^{\otimes n}_\theta]{d} \mathcal{N}\left(-\frac{1}{2}h^\mathsf{T} I_\theta h, h^\mathsf{T} I_\theta h\right)$. Since $\mathrm{Cov}(I_\theta^{-1} \nabla_\theta l_n(\theta), h^\mathsf{T} \nabla_\theta l_n(\theta)) = h$,

$$\left(\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right), \log \frac{dp^{\otimes n}_{\theta+h/\sqrt{n}}}{dp^{\otimes n}_{\theta}}\right) \xrightarrow[p^{\otimes n}_\theta]{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ -\frac{1}{2}h^\mathsf{T} I_\theta h \end{bmatrix}, \begin{bmatrix} I_\theta^{-1} & h \\ h^\mathsf{T} & h^\mathsf{T} I_\theta h \end{bmatrix}\right).$$

12.1.3 Step 3: Characterise L by Le Cam's third lemma

Theorem 12.2 (Le Cam's third lemma)**.** *Let P_n and Q_n be sequences of probability measures and X_n a sequence of random vectors. If*

$$\left(X_n, \log \frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} \mathcal{N}\left(\begin{bmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{bmatrix}, \begin{bmatrix} \Sigma & \tau \\ \tau^\mathsf{T} & \sigma^2 \end{bmatrix}\right),$$

then $X_n \xrightarrow[Q_n]{d} \mathcal{N}(\mu + \tau, \Sigma)$.

τ is generally a function of h, but not necessarily equal to h (as in the case of the MLE).

Observe that we have assumed a relationship between the mean and variance of V. When we have this relationship, the term σ^2 vanishes in the distribution of L. So $X_n \xrightarrow[p^{\otimes n}_\theta]{d} X$ and $X_n \xrightarrow[p^{\otimes n}_{\theta+h/\sqrt{n}}]{d} L$ with the variance of L completely determined by the variance of X.

Applying Le Cam's third lemma to the MLE, we have

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right) \xrightarrow[p^{\otimes n}_{\theta+h/\sqrt{n}}]{d} \mathcal{N}\left(h, I_\theta^{-1}\right),$$

or

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \left[\theta + \frac{h}{\sqrt{n}}\right]\right) \xrightarrow[p^{\otimes n}_{\theta+h/\sqrt{n}}]{d} \mathcal{N}\left(0, I_\theta^{-1}\right). \tag{31}$$

This is basically the classic (non-local) asymptotic MLE result. So it turns out perturbing by $h/\sqrt{n}$ doesn't change much. It also suggests that $\frac{1}{\sqrt{n}}$ is the right scling for the perturbation – it's likely a larger power would result in X_n diverging to infinity under $p_{\theta+h/\sqrt{n}}^{\otimes n}$ and a smaller power would result in convergence to a degenerate distribution.

Proof sketch of 12.2. See Question 1, Assignment 3 for a full proof. Use the characteristic function $\psi_L(t) = \mathbb{E}\left[e^{it^\intercal L}\right]$: We know that $\mathbb{P}\left[L \in B\right] = \mathbb{E}\left[\mathbb{1}\{X \in B\}\, e^V\right]$. By InSiPoD, this implies,

$$\int f(l)dP_L(l) = \mathbb{E}\left[f(X)e^V\right],$$

for every measurable function f. Hence

$$\psi_L(t) = \mathbb{E}\left[e^{-t^\intercal X}e^V\right].$$

Observe that the RHS is the characteristic function of (X, V) at $(t, -i)$. We know the characteristic function of (X, V) since (X, V) is multivariate normal. With some simple algebra we can then show that $\psi_L(t)$ has the required form. □

12.2 Local asymptotic analysis for other estimators

Above we found the local asymptotic distribution for the MLE by working through the three steps above. For other estimators, we can also go through these steps again to determine their local asymptotic distribution. But this would be on a case-by-case basis. Can we get a general set of results for a number of estimators? For what estimators Δ_n and what kinds of parametric families $\{p_\theta\}$ can we characterise the joint distribution of Δ_n and $\frac{dp_{\theta+h/\sqrt{n}}^{\otimes n}}{dp_\theta^{\otimes n}}$ under the local alternative? (Note that we will always have contiguity of the local alternative distribution under some mild smoothless conditions.)

The vague answer is that we can do this for any parametric family that asymptotically it locally looks Gaussian. We will formalise this answer.

12.2.1 Local asymptotic normality (LAN)

Definition 12.3. A class of models $\{p_\theta : \theta \in \mathcal{H}\}$ with $\mathcal{H} \subset \mathbb{R}^d$ is *locally asymptotically Normal* (LAN) at $\theta_0 \in \mathcal{H}$ with *precision* (aka information) $K \in \mathbb{R}^{d\times d}$ if there exists

(i) invertible matrices $r_n \in \mathbb{R}^{d\times d}$;

(ii) random vectors Δ_n with

$$\Delta_n \xrightarrow[p_{\theta_0}^{\otimes n}]{d} \mathcal{N}(0, K);$$

such that for every sequence $h_n \to h$,

$$\log \frac{dp^{\otimes n}_{\theta_0 + r_n^{-1} h_n}}{dp^{\otimes n}_{\theta_0}} = h^\intercal \Delta_n - \frac{1}{2} h^\intercal K h + o_{p_{\theta_0}^{\otimes n}}(\|h\|).$$

In the earlier calculations, we were relying on the score and the asymptotic normality of the score with variance I_θ. In the definition of LAN, we replace the score function with Δ_n and the Fisher information with K. So we can do the exact same calculations using $Z_n = K^{-1}\Delta_n + o_{p_{\theta_0}^{\otimes n}}(1)$ instead of

$$X_n = \sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right) = I_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta l_n(\theta) + o_{p_{\theta_0}^{\otimes n}}(1),$$

and get

$$\left(Z_n, \log \frac{dp^{\otimes n}_{\theta_0 + r_n^{-1} h_n}}{dp^{\otimes n}_{\theta_0}}\right) \xrightarrow[p_{\theta_0}^{\otimes n}]{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ -\frac{1}{2} h^\intercal K h \end{bmatrix}, \begin{bmatrix} K^{-1} & h \\ h^\intercal & h^\intercal K h \end{bmatrix}\right).$$

Then Le Cam's third lemma gives

$$Z_n \xrightarrow[p^{\otimes n}_{\theta_0 + r_n^{-1} h_n}]{d} \mathcal{N}(h, K^{-1}).$$

So we get local asymptotic analysis for a large class of families. But Z_n is a very specific estimator, since it is defined by K and Δ_n, which are specified by the definition of LAN. What about other types of estimators? With the LAN property, we can also derive results for M- and Z-estimators. For Z-estimators, we have

$$\sqrt{n}\left(T_n - \theta\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_\theta(X_i) + o_{p_{\theta_0}^{\otimes n}}(1),$$

(see Theorem 7.1), which has a Gaussian limit distribution by the CLT. Since

$$\left(T_n, \log \frac{dp^{\otimes n}_{\theta_0 + r_n^{-1} h_n}}{dp^{\otimes n}_{\theta_0}}\right),$$

is asymptotically multivariate Normal, all we need to get its joint asymptotic distribution is to calculate

$$\operatorname{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\theta(X_i), \Delta_n\right).$$

Hence, we can do local asymptotic analysis when the model is LAN and we can calculate this covariance.

Example 12.4. Examples of LAN families:

1. In Theorem 12.1 (which showed that $\log \frac{dp^{\otimes n}_{\theta_0+h/\sqrt{n}}}{dp^{\otimes n}_{\theta_0}}$ converges in QMD families), we used $r_n = \sqrt{n}I$, $\Delta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta(\theta_0, X_i)$ and $K = I_{\theta_0}$. Hence any family that is QMD at θ_0 is LAN at θ_0, as long as the parameter space $\mathcal{H}$ is open.
2. Gaussian local families: $Y_i = \frac{h_n}{\sqrt{n}} + \xi_i$ where $\xi_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$.
3. See [vdV] and the Homework for more examples.

12.2.2 Local asymptotic analysis for non-iid data

Up until now, we have consider iid data and the n-fold product. Yet this has not actually been necessary for our conclusions. In the non-iid scenario, we can generalise LAN as follows:

Definition 12.5. A sequence (indexed by n) of model families $\{p_{\theta,n} : \theta \in \mathcal{H}, n \in \mathbb{N}\}$ is *locally asymptotically Normal* at $\theta_0 \in \mathcal{H}$ with *precision* $K \in \mathbb{R}^{d\times d}$ if there exists

(i) invertible matrices $r_n \in \mathbb{R}^{d\times d}$;

(ii) random vectors Δ_n with

$$\Delta_n \xrightarrow[p_{\theta,n}]{d} \mathcal{N}(0, K);$$

such that for every sequence $h_n \to h$,

$$\log \frac{dp_{\theta_0 + r_n^{-1} h_n, n}}{dp_{\theta,n}} = h^\mathsf{T} \Delta_n - \frac{1}{2} h^\mathsf{T} K h + o_{p_{\theta,n}}(\|h\|) . \tag{32}$$

In this setup we observe data $(X_1, \ldots, X_n) \sim p_{\theta,n}$ (not necessarily iid). As long as we have the expansion of the log likelihood ratio as in (32), we can do all of the computations as in section 12.2.1 and obtain the same conclusions.

This is useful for time series or other autoregressive data – for example when $X_t = \theta X_{t-1} + Z_t$.

12.3 Recap of the last four lectures

We began this topic four lectures ago by studying the Hodges estimator through a risk analysis. This motivated the need for a local asymptotic analysis of risk. We made a jump to local asymptotic distribution since a distributional analysis gives more refined information as compared to the risk. We found that we needed a notion of asymptotic absolute continuity, which we formalised in the definition of contiguity. Today we looked at how to calculate the local asymptotic distribution. But recall why we began this local analysis: we wanted to compare estimators with the goal of choosing the 'optimal' estimator. In the next class, we will develop a notion of asymptotic optimality so that we can do this.

13 Lecture 9/3

Overview for today's lecture: We will discuss notions of asympotic optimality of estimators: 1) Asymptotic relative efficiency (ARE); 2) what is an optimal minimax estimator in terms of local asymptotic risk; 3) what are other sensible notions of optimality?

13.1 Asymptotic relative efficiency (ARE)

The basic (non-asymptotic) idea here is to compare the variances of two (or finitely many) proposal estimators. The optimal estimator will be the one with minimum variance. This analysis only makes sense if the proposed estimators are unbiased.

For asymptotic analysis, we consider estimators which are consistent – or which are centred around the parameter of interest (after appropriate scaling) – and we compare their asymptotic variances.

Definition 13.1. Suppose two sequences of estimators $\{\delta_n\}, \{\delta_n'\}$ satisfy

$$\begin{aligned}\sqrt{n}\left(\delta_n - g(\theta)\right) &\xrightarrow[\theta]{d} \mathcal{N}\left(0, \sigma_1^2(\theta)\right)\\ \sqrt{n}\left(\delta_n' - g(\theta)\right) &\xrightarrow[\theta]{d} \mathcal{N}\left(0, \sigma_2^2(\theta)\right),\end{aligned}$$

where $g(\theta)$ is the parameter of interest. Then the *asymptotic relative efficiency* of δ_n with respect to δ_n' is

$$\text{ARE} = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Note that we assumed the estimators are asymptotically Gaussian for simplicity but this is not necessary – we only need to assume the estimators converge to a distribution with a variance.

If ARE > 1 then we prefer δ_n. With multiple estimators that are all centred around $g(\theta)$, pick the one with the minimum asymptotic variance.

This definition of optimality is very restrictive, since we must assume the limit distribution's variance characterises the spread of the estimator; and that ranking of the variances across the two (or more) limit distributions corresponds to a ranking on the spreads of the estimators. In general, to compare between multiple estimators, just looking at the variance may not suffice. We should compare risks with respect to a loss function that reflects the desired properties of the estimator for the particular application in mind.

13.2 Minimaxity for local asymptotic risk

Setup: Suppose δ_n is $\sqrt{n}$-consistent for θ and consider the squared error loss. The asymptotic risk is

$$R_s(\delta_n, \theta) = \lim_{n\to\infty} n\mathbb{E}\left[\delta_n - \theta\right]^2.$$

This depends on the (unknown) value of θ. How should we summarise $R_s(\delta_n, \theta)$ into a single number for each estimator δ_n. Minimaxity takes the worst case risk $\sup_\theta R_s(\delta_n, \theta)$.

Can we generalise minimaxity to asymptotic risk? No, at least not straight away due to the problems highlighted by the Hodge's estimator. Recall the asymptotic risk of the Hodge's estimator

$$R_s\left(\hat{\theta}_n^{\mathrm{H}}, \theta\right) = \begin{cases} 1 & \text{if } \theta \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

So $\hat{\theta}_n^{\mathrm{H}}$ will be asymptotically minimax if the MLE is, since they both have the same worse case limit risk. But this doesn't capture the finite sample behaviour, nor does it capture the fact that $\lim_{n\to\infty} \sup_\theta R_n\left(\hat{\theta}_n^{\mathrm{H}}, \theta\right) = \infty$ (26).

The fix is to apply minimaxity to the local asymptotic risk. Define

$$R_s(\delta_n, \theta, h) = \lim_{n\to\infty} n\mathbb{E}\left[\delta_n - \theta - \frac{h}{\sqrt{n}}\right]^2,$$

where the expectation is with respect to the local alternative $p_{\theta+h/\sqrt{n}}^{\otimes n}$. (Note that we have fixed the scaling at $\frac{1}{\sqrt{n}}$ since we decided in Section 4 that this is the right scale for studying local asymptotic risk.)

The optimal minimax estimator δ_n in the sense of the local asymptotic risk is defined as the minimiser of

$$\sup_h R_s(\delta_n, \theta, h), \tag{33}$$

for every θ (assuming that such a minimiser exists and is unique).

This seems like a very strong notion. It is not trivial that such an optimal estimator would exist. Yet we will see that one does, assuming the parametric family has some smoothness conditions.

We've seen how to calculate the local asymptotic distribution. But now we need to be able to calculate the local asymptotic risk for a large class of estimators simultaneously – otherwise we will have no hope of being able to find the optimal estimator. Thankfully, there is a way to simplify this problem considerably.

Theorem 13.2 (Theorem 8.3 of [vdV]). *Let $\{p_\theta : \theta \in \mathcal{H}\}$ be QMD at $\theta_0 \in \mathrm{Int}\mathcal{H}$, with non-singular Fisher information I_{θ_0}. Let g be differentiable at θ_0 and δ_n be an estimator of $g(\theta_0)$. Suppose that*

$$\sqrt{n}\left[\delta_n - g\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \xrightarrow[p_{\theta_0+h/\sqrt{n}}^{\otimes n}]{d} L_{\theta_0,h}, \tag{34}$$

for all h. Then there exists a randomised function[**] *T of $X \sim \mathcal{N}\left(h, I_{\theta_0}^{-1}\right)$ such that*

$$T - Ah \sim L_{\theta_0,h},$$

for all h, where A is the Jacobian of g at θ_0.

We delay the proof of this Theorem until the end of the lecture.

The point of this Theorem is that it allows us to characterise $L_{\theta_0,h}$ for all h. (34) is just saying that δ_n (appropriately centred and scaled) converges weakly under the local alternative, for all h. Usually, we know that (34) holds by using Theorems 11.1 and (12.1) – so it is not as strong an assumption as it may appear – yet until now we haven't had a general way to characterise $L_{\theta_0,h}$.

Note that X is not 'fresh', in the sense that it is not independent of everything else. In fact, it turns out that X depends on the local likelihood ratio.

Insights from this theorem:

1. $L_{\theta_0,h}$ is essentially a function of a single observation X from a Gaussian. So instead of trying to work with $L_{\theta_0,h}$, we can do all our calculations with $T - Ah$ under the Gaussian model.

2. For all QMD families, the local asymptotic distribution of any estimator sequence is completely described by the distribution of $T - (\frac{\partial g}{\partial \theta})_{ij}h$ where T is a function of $X \sim \mathcal{N}\left(h, I_{\theta_0}^{-1}\right)$.

3. The downside is that we don't know how T transforms X, so it is harder to go further than this.

[**] A function is randomised if it depends on a uniform random variable which is independent of everything else.

Assume that $\{p_\theta\}$ is QMD at all θ and that (34) can be strengthened to

$$\mathbb{E}\left(\sqrt{n}\left[\delta_n - g\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right]\right)^2 \to \mathbb{E}L^2_{\theta_0,h}, \tag{35}$$

(where the square is component-wise if δ_n, g and $L_{\theta_0,h}$ are multi-dimensional). Also suppose that $g(\theta) = \theta$. Then

$$R_s(\delta_n, \theta, h) = \mathbb{E}\|T - h\|^2,$$

since $A = I$, where $T = f(X)$ with $X \sim \mathcal{N}\left(h, I_{\theta_0}^{-1}\right)$ and $\|\cdot\|$ is the Euclidean norm. So (33) is equivalent to minimising

$$sup_h \mathbb{E}\|T - h\|^2.$$

Considering T as an estimator of the location of $\mathcal{N}\left(h, I_{\theta_0}^{-1}\right)$, we have transformed the calculation of local asymptotic risk minimaxity to the problem of minimaxing the squared error loss in a Gaussian location family. We know that the MLE of h is minimax with respect to square error loss, so we would expect the MLE of θ to be optimal in the sense of (33). The next theorem verifies this intuition.

Theorem 13.3. *Let p_θ be QMD at $\theta_0 \in \mathrm{Int}\mathcal{H}$ with non-singular Fisher information I_{θ_0}. Let g be differentiable at θ_0 and δ_n an estimator of $g(\theta_0)$. Suppose that*

$$\sqrt{n}\left[\delta_n - g\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \xrightarrow[p^{\otimes n}_{\theta_0+h/\sqrt{n}}]{d} L_{\theta_0,h},$$

for all h. (Note that these are exactly the assumptions of the previous theorem.) Then

$$\sup_h R_s(\delta_n, \theta_0, h) \geq I_{\theta_0}^{-1}.$$

We know that

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta_0 - \frac{h}{\sqrt{n}}\right) \xrightarrow[p^{\otimes n}_{\theta_0+h/\sqrt{n}}]{d} \mathcal{N}\left(0, I_{\theta_0}^{-1}\right),$$

when $\{p_\theta\}$ is QMD at θ_0. Assuming also that the second moments converge (as in (35)), then we get $R_s(\hat{\theta}^{\mathrm{ML}}, \theta_0, h) = I_{\theta_0}^{-1}$ for all h. Theorem 13.3 implies that the MLE is optimal in the sense of local asymptotic risk minimaxity (33), assuming that the family is QMD at all $\theta \in \mathcal{H}$ and $\mathcal{H}$ open.

Proof. We will prove the Theorem for $g(\theta) = \theta$. By Theorem 13.2,

$$\sqrt{n}\left(\delta_n - \theta - \frac{h}{\sqrt{n}}\right) \xrightarrow{d} T - h.$$

Then

$$\begin{aligned}
\sup_h R_s(\delta_n, \theta, h) &= \sup_h \lim_{n\to\infty} n\mathbb{E}\left(\delta_n - \theta - \frac{h}{\sqrt{n}}\right)^2 \\
&= \sup_h \liminf_{n\to\infty} n\mathbb{E}\left(\delta_n - \theta - \frac{h}{\sqrt{n}}\right)^2 \\
&\geq \sup_h \mathbb{E}\left(T - h\right)^2 \\
&\geq \inf_T \sup_h \mathbb{E}\left(T - h\right)^2 \\
&= I_{\theta_0}^{-1},
\end{aligned}$$

where the third line follows by the Portmanteau theorem using the non-negative continuous function $f(x) = x^2$; the $\inf_T$ in the second last line is over all possible functions T of $X \sim \mathcal{N}\left(h, I_{\theta_0}^{-1}\right)$; and the final line follows since we know from Stat211, that $T = \hat{h}^{\mathrm{ML}} = X$ is minimax with respect to squared error loss in a Gaussian location family, and X has variance $I_{\theta_0}^{-1}$. □

Proof of Theorem 13.2. Our goal is to show there exists a randomised function T such that $T - Ah = L_{\theta_0,h}$ for all h. Write out

$$\sqrt{n}\left[\delta_n - g\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] = \sqrt{n}\left[\delta_n - g\left(\theta_0\right)\right] - \sqrt{n}\left[g\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - g\left(\theta_0\right)\right]. \quad (36)$$

We have already seen that the second term goes to Ah in probability (see section 9.3) since g is differentiable at θ_0. We need to understand the first term $X_n = \sqrt{n}\left[\delta_n - g\left(\theta_0\right)\right]$ under the local alternative distribution. By Le Cam's main theorem (Theorem 11.1), we can determine this by studying

$$V_n = \left(X_n, \log\frac{dp_{\theta_0+h/\sqrt{n}}^{\otimes n}}{dp_{\theta_0}^{\otimes n}}\right),$$

under $p_{\theta_0}^{\otimes n}$. We know that $X_n \xrightarrow[p_{\theta_0}^{\otimes n}]{d} L_{\theta_0,0}$ by assumption and that

$$\log \frac{dp_{\theta_0+h/\sqrt{n}}^{\otimes n}}{dp_{\theta_0}^{\otimes n}} = h^\mathsf{T} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_\theta(X_i) - \frac{1}{2} h^\mathsf{T} I_{\theta_0} h + o_{p_{\theta_0}^{\otimes n}}(1),$$

since we assume QMD.

So $V_n = O_{p_{\theta_0}^{\otimes n}}(1)$. Prokhorov's theorem states that there exists a subsequence $\{n_j\}$ such that V_n has a weak limit along $\{n_j\}$. That is,

$$V_{n_j} = \left(X_{n_j}, \log \frac{dp_{\theta_0+h/\sqrt{n_j}}^{\otimes n_j}}{dp_{\theta_0}^{\otimes n_j}} \right) \xrightarrow[p_{\theta_0}^{\otimes n_j}]{d} (S, h^\mathsf{T} I_{\theta_0} \Delta - \frac{1}{2} h^\mathsf{T} I_{\theta_0} h),$$

as $j \to \infty$, where $\Delta \sim \mathcal{N}(0, I_{\theta_0}^{-1})$ is the limiting distribution of $h^\mathsf{T} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_\theta(X_i)$; and S is some random variable.

We need to express S as some function of Δ. We claim that

$$S = T(\Delta, U), \tag{37}$$

where U is a uniform random variable independent of Δ. Why? We will prove it for the univariate case where $X_n, S \in \mathbb{R}$. For a given value δ of Δ, let $F_{S|\Delta=\delta}(s)$ be the conditional distribution of S. Then by the probability integral transform (PIT),

$$F_{S|\Delta=\delta}^{-1}(U) \sim S,$$

and $T = F_{S|\Delta=\delta}^{-1}$ is a function of Δ and U as desired. (This naive hack of using the conditional distribution and PIT is surprisingly useful.)

By Theorem 11.1,

$$X_{n_j} \xrightarrow[p_{\theta_0+\frac{h}{\sqrt{n_j}}}^{\otimes n_j}]{d} L,$$

with L given by

$$\begin{aligned}
\mathbb{P}(L \in B) &= \mathbb{E}\left[\mathbb{1}\{T(\Delta, U) \in B\} \exp\left(h^\mathsf{T} I_{\theta_0} \Delta - \frac{1}{2} h^\mathsf{T} I_{\theta_0} h \right) \right] \\
&= \mathbb{E}_U \int \mathbb{1}\{T(\Delta, U) \in B\} \exp\left(h^\mathsf{T} I_{\theta_0} \Delta - \frac{1}{2} h^\mathsf{T} I_{\theta_0} h \right) \frac{\exp\left(-\frac{1}{2}\delta^\mathsf{T} I_{\theta_0} \delta\right)}{\sqrt{2\pi I_{\theta_0}^{-1}}} d\delta
\end{aligned}$$

$$= \mathbb{E}_U \int \mathbb{1}\{T(\Delta, U) \in B\} \frac{1}{\sqrt{2\pi I_{\theta_0}^{-1}}} \exp\left(-\frac{1}{2}(\delta - h)^\intercal I_{\theta_0}(\delta - h)\right) d\delta$$

$$= \mathbb{P}\big(T(\Delta, U) \in B | U \text{ uniform independent of } \Delta \sim \mathcal{N}(h, I_{\theta_0}^{-1})\big),$$

where the second line follows by explicitly writing out the expectation with respect to Δ, noting that $f(\delta) = \frac{1}{\sqrt{2\pi I_{\theta_0}^{-1}}} \exp\left(-\frac{1}{2}\delta^\intercal I_{\theta_0}\delta\right)$ is the density of Δ. Thus $L \sim T(\Delta, U)$ and by (36),

$$\sqrt{n_j}\left[\delta_{n_j} - g\left(\theta_0 + \frac{h}{\sqrt{n_j}}\right)\right] = X_{n_j} - Ah + o(1) \xrightarrow[p_{\theta_0 + \frac{h}{\sqrt{n_j}}}^{\otimes n_j}]{d} T(\Delta, U) - Ah.$$

Hence we have established the required result along a subsequence n_j. Extending this result to hold on the entire sequence will complete the proof. Yet we assumed the entire sequence weakly converges, and its limit must equal the limit on the subsequence $\{n_j\}$. □

14 Lecture 11/3

14.1 Regular estimators

Definition 14.1. An estimator δ_n is *regular* at θ_0 (for estimating $g(\theta_0)$) if, for all h,

$$\sqrt{n}\left[\delta_n - g\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \xrightarrow[p_{\theta_0 + h/\sqrt{n}}^{\otimes n}]{d} L_{\theta_0}, \tag{38}$$

where L_{θ_0} does not depend on h.

So an estimator is regular if its local asymptotic distribution doesn't depend on the direction h it approaches θ_0.

If an estimator is regular, then small vanishing changes (i.e. the direction h) around θ_0 does not affect the limit distribution. So regular estimators converge to their limit in a "locally uniform" fashion.

Example 14.2.

1. The MLE is regular since we showed that $L_{\theta_0} \sim \mathcal{N}(0, I_{\theta_0}^{-1})$ (see 31).
2. The Hodges estimator

$$\hat{\theta}_n^{\mathrm{H}} = \begin{cases} \bar{X}_n & \text{if } \left|\bar{X}_n\right| > n^{-1/4}, \\ \epsilon \bar{X}_n & \text{otherwise,} \end{cases}$$

   is not regular since

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{H}} - \frac{h}{\sqrt{n}}\right) \xrightarrow[p_{h/\sqrt{n}}^{\otimes n}]{d} \mathcal{N}\left(h(1-\epsilon), \epsilon^2\right).$$

   (See Homework 2.)

The following theorem establishes a notion of the "best" possible asymptotic distribution.

Theorem 14.3 (Theorem 8.8 of [vdV])**.** *Let p_θ be QMD at $\theta_0 \in \mathrm{Int}\mathcal{H}$ with non-singular Fisher information I_{θ_0}. Let g be differentiable at θ_0. Then the local asymptotic limit distribution L_{θ_0} (from 38) for any regular estimator δ_n of $g(\theta_0)$ satisfies*

$$L_{\theta_0} = Z_{\theta_0} + \Delta_{\theta_0},$$

where $Z_{\theta_0} \sim \mathcal{N}\left(0, A I_{\theta_0}^{-1} A\right)$; A is the Jacobian of g at θ_0; and Δ_{θ_0} is a random variable independent of Z_{θ_0} (but beyond that we cannot say more about Δ_{θ_0}).

So $\mathrm{Var} L_{\theta_0} = \mathrm{Var} Z_{\theta_0} + \mathrm{Var}\Delta_{\theta_0} \geq A I_{\theta_0}^{-1} A$ – that is, the local asymptotic variance of a regular estimator is always bounded below by $A I_{\theta_0}^{-1} A$. Thus, a regular estimator is asymptotically efficient if its asymptotic variance is $A I_{\theta_0}^{-1} A$.

Recall (31):

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \left[\theta + \frac{h}{\sqrt{n}}\right]\right) \xrightarrow[p_{\theta+h/\sqrt{n}}^{\otimes n}]{d} \mathcal{N}\left(0, I_\theta^{-1}\right).$$

By the delta method (assuming A is non-singular),

$$\sqrt{n}\left(g\left(\hat{\theta}^{\mathrm{ML}}\right) - g\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \xrightarrow[p_{\theta+h/\sqrt{n}}^{\otimes n}]{d} \mathcal{N}\left(0, A I_\theta^{-1} A\right),$$

and $g\left(\hat{\theta}^{\mathrm{ML}}\right)$ is the MLE of $g(\theta_0)$ by equivariance. So the MLE achieves optimal local asymptotic variance (assuming that A is non-singular).

14.1.1 Superefficient Estimators

Question: Theorems 13.2 and 14.3 both formalise the notion that the MLE is optimal in 'nice' families. But then how do superefficient estimators[††] – such as the Hodges estimator – exist (in these 'nice' families)? The following theorem reconciles this apparent contradiction.

Theorem 14.4 (Theorem 8.9 of [vdV]). *Suppose the model $\{p_\theta : \theta \in \mathcal{H}\}$ is QMD at every $\theta \in \mathcal{H}$ with non-singular Fisher information I_{θ_0}. Let $g : \mathcal{H} \to \mathbb{R}^k$ be a differentiable function and δ_n a sequence of estimators of $g(\theta)$ such that*

$$\sqrt{n}\,(\delta_n - g(\theta)) \xrightarrow[p_\theta^{\otimes n}]{d} L_\theta,$$

for all θ. (All this is saying is that δ_n converges in distribution.)

Then there exists a random vector Δ_θ such that, for Lebesgue-a.e. θ,

$$L_\theta = Z_\theta + \Delta_\theta,$$

where $Z_\theta \sim \mathcal{N}(0, A_\theta I_\theta^{-1} A_\theta)$ and A_θ is the Jacobian of g at θ. (We no longer conclude that Δ_θ is independent of Z_θ.)

Hence, superefficient estimators in QMD families that converge weakly for all θ must have variance at least $A_\theta I_\theta^{-1} A_\theta$ for almost all θ. That is, they can only be superefficient on a set of Lebesgue measure zero!

14.2 Local asymptotic minimax theorem

Theorem 14.5 (Theorem 8.11 of [vdV]). *Given a family that is QMD at $\theta_0 \in \mathcal{H} \subset \mathbb{R}^k$ with non-singular Fisher information I_{θ_0}, a function g differentiable at θ_0, an*

[††] An estimator is superefficient if its asymptotic variance is less than $I(\theta_0)^{-1}$. See Section 4, part 2 for more details.

estimator δ_n of $g(\theta_0)$ and a bowl-shaped loss function[‡‡] *l,*

$$\sup_I \liminf_{n\to\infty} \sup_{h\in I} \mathbb{E}_{p^{\otimes n}_{\theta_0+h/\sqrt{n}}} l\left(\sqrt{n}\left[\delta_n - g(\theta_0 + h/\sqrt{n})\right]\right) \geq \mathbb{E}_{X\sim p(\cdot)} l(X), \tag{39}$$

where $\sup_I$ *is taken over all finite subsets $I \subset \mathbb{R}^k$; $p(\cdot)$ is the density of $\mathcal{N}(0, AI_{\theta_0}^{-1}A)$; and A is the Jacobian of g at θ_0.*

$\mathbb{E}_{p^{\otimes n}_{\theta_0+h/\sqrt{n}}} l\left(\sqrt{n}\left[\delta_n - g(\theta_0 + h/\sqrt{n})\right]\right)$ is the local risk and the LHS of (39) is the worst case local risk. We use $\liminf_{n\to\infty}$ since we want to bound the worst case local risk from below.

14.3 Connections to modern research

This module on local asymptotics has presented classical results, but they are used in modern statistics. See [JO20, AKJ20, BM18, LS20] for papers on spike matrices $M = \lambda vv^{\mathsf{T}} + Z$. Notions of asymptotic optimality in high dimensional regression remain unresolved. The following are some open questions in this area:

1. What is the notion of asymptotically optimal estimators?

2. What kind of local expansions to the log likelihood ratio are useful and what do you get out of this?

3. What is the right local resolution? Just using $n^{-1/2}$ is naturally not enough, since p is also diverging. We will need to use something like p^α, n^β, but for what values of α, β?

[‡‡] A bowl shaped loss function l is a function with values in $[0,\infty]$ such that the sublevel sets $\{x : l(x) \leq c\}$ are convex and symmetric about the origin. The loss of δ_n estimating $g(\theta)$ is $l(\delta_n - g(\theta))$. The bowl-shaped property ensures that $l(\delta_n - g(\theta)) = l(g(\theta) - \delta_n)$ and that, roughly, the loss increases with $\|\delta_n - g(\theta)\|$.

15 Lecture 18/3

15.1 Hypothesis testing

In Stat211, we looked at testing in finite samples, the Neyman-Pearson framework and the UMP. We now move to an asymptotic study of hypothesis testing, since – analogous to estimation – asymptotics provides a unifying general theory.

15.1.1 Setup and preliminary definitions

Our goal is to test whether the null hypothesis $H_0 : \theta \in \Theta_0$ holds versus the alternative hypothesis $H_1 : \theta \in \Theta_1$. We observe $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$. Write T_n for the test statistic based on the n observed samples.

Definition 15.1. The *critical function* ϕ_n of a test is the statistic

$$\phi_n(X_1, \ldots, X_n) = \mathbb{P}\left(T_n \text{ rejects the null given observations } X_1, \ldots, X_n\right).$$

(The probability in ϕ_n solely reflects the fact that there may be some randomisation within the test T_n – it is not a probability over the X_i's or something else.) The critical function completely characterises the test; it will prove to be a convenient way of working with tests.

Definition 15.2. For a test T_n that rejects if T_n is in some critical region K_n, define the *power function*

$$\pi_n : \theta \mapsto \mathbb{P}_\theta(T_n \in K_n).$$

More generally,

$$\pi_n : \theta \mapsto \mathbb{E}_\theta \phi_n.$$

(The expectation/probability in π_n is taken over the data X_i.) The power function $\pi_n(\theta)$ is the probability of rejecting the null hypothesis, under parameter θ.

Definition 15.3. The *size* of a test is defined as

$$\sup_{\theta \in \Theta_0} \pi_n(\theta).$$

A test is *level* α if its size is no greater than α. A test is *asymptotically level* α if

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

Definition 15.4. Define the *limiting power function* of a test as the point-wise limit of the power

$$\pi(\theta) = \lim_{n\to\infty} \pi_n(\theta),$$

assuming that the limit exists (which it usually does).

A test with power function π_n is better than one with power function $\underline{\pi}_n$ if

$$\begin{aligned} \pi_n(\theta) &\le \underline{\pi}_n(\theta) \quad \forall\theta\in\Theta_0, \\ \pi_n(\theta) &\ge \underline{\pi}_n(\theta) \quad \forall\theta\in\Theta_1. \end{aligned} \tag{40}$$

Naively, we could make an analogous asymptotic definition using the limiting power function. However, the following example shows that such a definition will have very little utility.

Example 15.5. Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$. Suppose we want to test $H_0 : \theta = 0$ versus the two-sided alternative $H_1 : \theta \neq 0$. Let

$$T_n = \mathbb{1}\left\{ \left|\bar{X}_n\right| > \frac{z_{1-\alpha/2}}{\sqrt{n}} \right\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$. Since $\bar{X}_n \sim \mathcal{N}(\theta, \frac{1}{n})$,

$$\begin{aligned} \pi(\theta) &= \lim_{n\to\infty} \mathbb{P}_\theta\left(\left|\bar{X}_n\right| > \frac{z_{1-\alpha/2}}{\sqrt{n}} \right) \\ &= \lim_{n\to\infty} 1 - \mathbb{P}_\theta\left(z_{-\alpha/2} - \sqrt{n}\theta < Z < z_{1-\alpha/2} - \sqrt{n}\theta \right) \\ &= \begin{cases} 1 & \text{if } \theta \neq 0, \\ \alpha & \text{if } \theta = 0, \end{cases} \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. We can take α to be very small and then we can almost perfectly distinguish between the null and the alternative asymptotically. This phenomenon – the power going to 1 and α – is common. Hence the asymptotic version of (40) will not be a useful way to compare tests. We need a better way to distinguish tests asymptotically.

15.1.2 The sign test

Consider the scenario where $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ and $X_1, \dots, X_n$ are iid from a location family $F(x - \theta)$. Assume that $F(x)$ has a unique median 0. The sign statistic is defined as

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i > 0\}.$$

Define

$$\mu(\theta) = \mathbb{E}_\theta S_n = \mathbb{P}_\theta(X_i > 0) = 1 - F(-\theta).$$

Further

$$\sigma^2(\theta)\mathrm{Var}(S_n) = \frac{1}{n}\left[1 - F(-\theta)\right] F(-\theta),$$

since $nS_n \sim \mathrm{Binom}(n, 1 - F(-\theta))$. Then

$$\sqrt{n}(S_n - \mu(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)).$$

Under the null $\mu(0) = 1 - F(0) = \frac{1}{2}$ and $\mu^2(0) = \frac{1}{4}$. Hence the test that rejects when

$$\sqrt{n}\left(S_n - \frac{1}{2}\right) > \frac{1}{2} z_{1-\alpha},$$

has asymptotic size α. The power function can be calculated by

$$\begin{aligned}
\pi_n(\theta) &= \mathbb{P}_\theta\left[\sqrt{n}\left(S_n - \mu(0)\right) > \frac{1}{2} z_{1-\alpha}\right] \\
&= \mathbb{P}_\theta\left[\sqrt{n}\frac{S_n - \mu(\theta)}{\sigma(\theta)} > \frac{\frac{1}{2} z_{1-\alpha} - \sqrt{n}\left(\mu(\theta) - \mu(0)\right)}{\sigma(\theta)}\right] \\
&= 1 - \Phi\left[\frac{\frac{1}{2} z_{1-\alpha} - \sqrt{n}\left(\mu(\theta) - \mu(0)\right)}{\sigma(\theta)}\right] + o_{\mathbb{P}}(1).
\end{aligned}$$

We know that

$$\sqrt{n}\left(\mu(\theta) - \mu(0)\right) = \sqrt{n}\left(F(0) - F(-\theta)\right).$$

If $\theta > 0$ then $F(0) - F(-\theta)) > 0$ since the median is unique. Hence the argument of Φ goes to negative infinite and $\pi_n(\theta) \to 1$ under the alternative. On the other hand, when $\theta = 0$,

$$\pi_n(\theta) = 1 - \Phi\left(\frac{\frac{1}{2} z_{1-\alpha}}{\frac{1}{2}}\right) + o_{\mathbb{P}}(1) = \alpha + o_{\mathbb{P}}(1).$$

The limiting power function is thus

$$\pi(\theta) = \begin{cases} 1 & \text{if } \theta > 0, \\ \alpha & \text{if } \theta = 0. \end{cases}$$

This is another example of the phenomenon of asymptotic power going to one. We will define a test to be consistent if it observes this phenomenon.

15.1.3 Consistency (of hypothesis tests)

Definition 15.6. A sequence of tests with power function $\pi_n(\theta)$ is *asymptotically consistent* at level α if

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} \pi_n(\theta) \leq \alpha, \tag{41}$$

and $\lim_{n\to\infty} \pi_n(\theta) = 1$, for all $\theta \in \Theta_1$.

Like consistency for estimators, we that all reasonable hypothesis test are consistent. It can be viewed as a minimum baseline for a 'decent' hypothesis test.

(41) is a departure from the Neyman-Pearson paradigm where we require level α for every finite n. Now we only care about the level in the limit. (41) is usually the easier of the two conditions of consistency to satisfy, since we can choose the critical region so as to bound the type 1 error. The second condition – a pointwise limiting power of 1 – is the main feature of consistency.

15.1.4 Local limiting power discussion

How can we compare between different tests asymptotically? Consistency is too easier to obtain, since the pointwise limit function is not really informative. To make an real comparison between sequences of consistent tests, we need to study the performance of tests in problems that get harder and harder as n increases.

Consistency shows that it is too easy to distinguish between a fixed null and a fixed alternative. One way to make the testing problem harder is to choose the null and and alternative hypothesis to move closer to each other as $n \to \infty$. The idea is very similar to the local asymptotic analysis that we have been already been considering:

fix the null and consider a sequence of alternative hypotheses that converge to the null. Then calculate the power of the test along this sequence of alternatives.

Recall the sign test. Suppose that $\theta_n \downarrow 0$. We know that

$$\pi_n(\theta_n) = 1 - \Phi\left(\frac{\frac{1}{2}z_{1-\alpha} - \sqrt{n}\,[F(0) - F(-\theta_n)]}{\sigma(\theta_n)}\right) + o_{\mathbb{P}}(1).$$

The limiting behaviour of $\pi_n(\theta_n)$ depends on the rate θ_n converges to zero and how this affects the rate that $F(0) - F(-\theta_n)$ converges to zero. Assuming that $\sigma(\cdot)$ is continuous at 0, there are two cases:

1. $\theta_n \to 0$ too fast so that

   $$\sqrt{n}\,[F(0) - F(-\theta_n)] \to 0,$$

   and hence $\pi_n(\theta_n) \to \alpha$. In this case, the testing problem is too hard, since we are unable to distinguish between the alternative and null hypotheses.

2. $\theta_n \to 0$ too slow so that

   $$\sqrt{n}\,[F(0) - F(-\theta_n)] \to \infty,$$

   and hence $\pi_n(\theta_n) \to 1$. In this case, the testing problem is too easy.

What is crucial is the rate of convergence of $F(0) - F(-\theta_n)$ as compared to the rate $\sqrt{n}$. What is the rate at which we require $\theta_n \to 0$ so that the testing problem is neither too easy nor too hard – ie so that $\lim_{n\to\infty} \pi_n(\theta_n) \in (\alpha, 1)$?

Assume that F is differentiable at zero with positive derivative. Then the Taylor expansion

$$\sqrt{n}\,[F(0) - F(-\theta_n)] = \sqrt{n}\theta_n f(0) + \sqrt{n}o(\theta_n). \tag{42}$$

If $f(0)$ is bounded then we need $\theta_n = \Theta(n^{-1/2})$, in order that (42) does not converge to zero or infinity. That is, θ_n should be converge to zero at rate $n^{-1/2}$. This is a specific example for the sign test, but the phenomenon is universal across a wide class of smooth parametric families. This is tied to why $\alpha = \frac{1}{2}$ is the right rate of convergence for local asymptotics.

For $\theta_n = \frac{h}{\sqrt{n}}$,

$$\pi_n\left(\frac{h}{\sqrt{n}}\right) \to 1 - \Phi\left(z_{1-\alpha} - 2hf(0)\right).$$

15.1.5 Local limiting power function

To recap the discussion of the previous section: in order to asymptotically compare two sequences of consistent tests for $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, consider the *local limiting power function*, defined as

$$\pi(h) = \lim_{n\to\infty} \pi_n \left(\frac{h}{\sqrt{n}}\right).$$

For a one sided test, consider $h > 0$; for a two sided test, consider any $h \in \mathbb{R}$; for a multi-dimensional parameter, consider any $h \in \mathbb{R}^k$.

For any sequence of "nice" test statistics, $\pi(h)$ has a "nice" formula; and this formula provides a way to compare consistent tests.

For simplicity, assume the one sided setting where $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Suppose that the sequence of tests reject the null for large values of T_n[§§], and for all $h \geq 0$,

$$\frac{\sqrt{n}\,(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \xrightarrow[p_{\theta_n}]{d} \mathcal{N}(0,1), \tag{43}$$

where $\theta_n = \frac{h}{\sqrt{n}}$ and $\mu(\theta_n), \sigma(\theta_n)$ are some know functions (commonly – but not necessarily – the mean and standard deviation of T_n).

(43) is what we mean by "nice" test statistic. (For a two sided test, T_n should reject for large absolute values and (43) should hold for all h.)

For the one sided test, (43) implies that the critical value $\frac{\sigma(0)z_{1-\alpha}}{\sqrt{n}} + \mu(0)$ defines an asymptotically level α test. Further,

$$\begin{aligned}\pi_n(\theta_n) &= \mathbb{P}_{\theta_n}\left[\sqrt{n}\,(T_n - \mu(0)) > \sigma(0)z_{1-\alpha}\right] \\ &= \mathbb{P}_{\theta_n}\left[\frac{\sqrt{n}\,(T_n - \mu(\theta_n))}{\sigma(\theta_n)} > \frac{\sigma(0)z_{1-\alpha} - \sqrt{n}\,(\mu(\theta_n) - \mu(0))}{\sigma(\theta_n)}\right].\end{aligned}$$

If $\theta_n = \frac{h}{\sqrt{n}}$, $\mu(\cdot)$ differentiable at zero and $\sigma(\cdot)$ continuous at zero, then under (43),

$$\pi\,(h) = 1 - \Phi\left[\frac{\sigma(0)z_{1-\alpha} - h\mu'(0)}{\sigma(0)}\right]$$

[§§] "The test statistic rejects the null for large values" is a common statement. It is a useful ambiguity since it allows us to specify the test without exactly specifying the rejection region. We like this since we want to be able to change the rejection region depending on the desired level α; we may also want to defer the specification of the rejection region until later.

$$= 1 - \Phi\left[z_{1-\alpha} - \frac{h\mu'(0)}{\sigma(0)}\right], \tag{44}$$

by similar derivations to the sign test example in the previous example. (44) is the "nice" formula for $\pi(h)$.

Therefore, to compare the asymptotic local power of tests which satisfy (43), it suffice to compare

$$\beta = \frac{\mu'(0)}{\sigma(0)}.$$

For one sided tests with $h > 0$, larger values of β result in larger asymptotic local power.

Theorem 14.7 of [vdV] summarises the discussion of this section.

16 Lecture 23/3

16.1 Slope and relative efficiency

Definition 16.1. Consider $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. Given a sequence of tests that reject the null for large values of T_n which satisfy (43),

$$\beta = \frac{\mu'(0)}{\sigma(0)},$$

is called the *slope* of the tests, assuming that μ is differentiable at zero and σ continuous at zero.

Now consider two sequences of tests that reject the null for large values of $T_{n,1}$ and $T_{n,2}$. Assume that they satisfy (43) and that μ_i is differentiable at zero, σ_i is continuous at zero and $\mu_i(0) > 0, \sigma_i(0) > 0$, for $i = 1, 2$. The *asymptotic relative efficiency (ARE)* of the tests is defined as the square of the ratio of the slopes:

$$\text{ARE} = \left(\frac{\beta_1}{\beta_2}\right)^2.$$

Typically, the ARE is studied under scenarios which satisfy the following assumptions:

1. The statistical model $\{p_{n,\theta} : \theta \geq 0\}$ satisfies a local smoothness condition around $\theta = 0$:
$$\|p_{n,\theta} - p_{n,0}\|_{\text{TV}} \to 0,$$
as $\theta \to 0$, for all n.

2. Both tests have asymptotic level α and pointwise power limiting to $\gamma \in (\alpha, 1]$ for all $\theta > 0$.

Under these assumptions, the ARE has a sample size interpretation – see the discussion around Theorem 13.9 of [vdV].

16.1.1 Limitations of the ARE

The ARE cannot handle nuisance parameters. For example, suppose we are interested in testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$ in a family $\{p_{n,\theta,\sigma} : \theta \geq 0\}$ parametrised by both θ and σ. (We argue that this is actually the more typical scenario than one without nuisance parameters.) In this setup, the ARE can depend on the value of σ. So for some σ, one test may be better, while for other σ, the other test may be better. Yet typically, we do not know the true σ, so we don't know which test to choose based on this ARE analysis. This motivates a better way to compare tests – asymptotic optimality of tests – which we will introduce next class.

16.2 Sufficient conditions for consistency

Lemma 16.2 (14.15 of [vdV]). *Let T_n be a sequence of statistics such that $T_n \xrightarrow{P_\theta} \mu(\theta)$ for every θ. Then the family of tests that reject the null hypothesis $H_0 : \theta = 0$ for large values*[¶¶] *of T_n is consistent against every θ with $\mu(\theta) > \mu(0)$.*

For a proof, see Section 8.

Example 16.3. Consider the two sample setting where we observe $X_1, \ldots, X_n \overset{iid}{\sim} F$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} G$. We want to test $H_0 : \mathbb{E}X = \mathbb{E}Y$. Consider the test statistic
$$\frac{\bar{Y} - \bar{X}}{S} \xrightarrow{P} \frac{\mathbb{E}\,(Y - X)}{\sigma},$$

[¶¶] So we are looking at one sided tests.

where S^2 is the sample variance of $Y_i - X_i$ and $\sigma^2 = \lim_{n\to\infty} \operatorname{Var}\left(\bar{Y} - \bar{X}\right)$. This lemma shows that the test is consistent against every alternative of the form $\mathbb{E}Y > \mathbb{E}X$.

Also see Lemma 14.16 in [vdV] (Lemma 2 in Section 8) for another set of sufficient conditions for consistency.

16.3 The Wald test

Consider the setup where $X \sim P_\theta$ (we could also have $X_1, \dots, X_n \overset{iid}{\sim} P_\theta$) and we want to test $H_0 : g(\theta) = 0$ against $H_1 : g(\theta) \neq 0$, where $g : \Theta \to \mathbb{R}^k$ is differentiable.

The Wald test statistic is based on $g(\hat{\theta}^{\mathrm{ML}})$. To control the Wald test's size, we need to understand $g(\hat{\theta}^{\mathrm{ML}})$ under the null. Assume regularity conditions so that

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right) \xrightarrow{d} \mathcal{N}(0, I_\theta^{-1}).$$

Then the delta method gives

$$\sqrt{n}\left(g(\hat{\theta}^{\mathrm{ML}}) - g(\theta)\right) \xrightarrow{d} \mathcal{N}(0, J_\theta I_\theta^{-1} J_\theta^{\mathsf{T}}),$$

where J_θ is the Jacobian of g at θ. Since $g(\theta) = 0$ under the null,

$$\sqrt{n} V_\theta^{-1/2} g(\hat{\theta}^{\mathrm{ML}}) \xrightarrow[H_0]{d} \mathcal{N}(0, I), \tag{45}$$

where $V_\theta = J_\theta I_\theta^{-1} J_\theta^{\mathsf{T}}$.

We would like to use (45) as our test statistic. There are two issues:

1. The null space Θ_0 may not be a singleton set. In this case, what θ should be used in V_θ in the 'test statistic' (45)?

2. Even if there is a unique θ in the null, $g(\theta) = 0$ may not yield a closed form expression for θ and hence it may be hard or impossible to compute V_θ.

So we cannot just plug in $V_\theta^{-1/2}$ for some $\theta \in \Theta_0$ into (45) and get a test statistic. Instead, assuming $\theta \mapsto V_\theta$ is a continuous function in θ, then

$$V_{\hat{\theta}^{\mathrm{ML}}} \xrightarrow{P_\theta} V_\theta,$$

and hence

$$\sqrt{n} V_{\hat{\theta}^{\mathrm{ML}}}^{-1/2} g(\hat{\theta}^{\mathrm{ML}}) \xrightarrow[H_0]{d} \mathcal{N}(0, I). \tag{46}$$

(46) is the definition of the Wald test. Equivalently, we can also use the test statistic

$$n g^{\mathsf{T}}(\hat{\theta}^{\mathrm{ML}}) V_{\hat{\theta}^{\mathrm{ML}}}^{-1} g(\hat{\theta}^{\mathrm{ML}}) \xrightarrow[H_0]{d} \chi_k^2,$$

where k is the dimension of $g(\theta)$.

We could replace $\hat{\theta}^{\mathrm{ML}}$ with any consistent and asymptotically Normal estimator. This is useful in practise when the MLE is difficult to compute, while another estimator (e.g. the one-step estimator) is not.

Questions: Is the Wald test consistent? Yes – this is not hard to prove using the fact

$$\sqrt{n} V_{\hat{\theta}^{\mathrm{ML}}}^{-1} \left(g(\hat{\theta}^{\mathrm{ML}}) - g(\theta) \right) \xrightarrow{d} \mathcal{N}(0, I).$$

What is the asymptotic local power of the Wald test? There are two subtleties: Firstly, we need to find the asymptotic distribution of $g(\hat{\theta}^{\mathrm{ML}})$ (with appropriate centring and scaling) under $p_{\theta+h/\sqrt{n}}^{\otimes n}$. We can't apply the delta method out of the box as the distribution is changing with n. (But I think it turns out that we can get a delta method for this case.) Also, we need $V_{\hat{\theta}^{\mathrm{ML}}}^{-1/2} \xrightarrow{P} V_\theta^{-1/2}$ under $p_{\theta+h/\sqrt{n}}^{\otimes n}$. We don't know yet know a result to handle this either; although proving one is easy:

Proposition 16.4. *If $\{X_n\}$ is a sequence of random variables such that $X_n \xrightarrow{P_n} X$ and $Q_n \triangleleft P_n$ then $X_n \xrightarrow{Q_n} X$ (where the convergences are in probability).*

The proof of this proposition is an easy consequence of contiguity.

With some calculations, we get

$$\sqrt{n} \left(g(\hat{\theta}^{\mathrm{ML}}) - g(\theta) \right) \xrightarrow[p_{\theta+h/\sqrt{n}}^{\otimes n}]{d} \mathcal{N}(J_\theta h, V_\theta),$$

and assuming contiguity – so that $V_{\hat{\theta}^{\mathrm{ML}}}^{-1/2} \xrightarrow[p_{\theta+h/\sqrt{n}}^{\otimes n}]{P} V_\theta^{-1/2}$ – we have

$$\sqrt{n} V_{\hat{\theta}^{\mathrm{ML}}}^{-1/2} g(\hat{\theta}^{\mathrm{ML}}) \xrightarrow[p_{\theta+h/\sqrt{n}}^{\otimes n}]{d} \mathcal{N}(V_\theta^{-1/2} J_\theta h, I).$$

16.3.1 The Wald test with nuisance parameters

As a running example for this section, consider $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are both unknown; and we are interested in testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.

More generally, we are in a setting where we only care about estimating functions of θ that depend on only a few co-ordinates of θ.

Notation: For $v \in \mathbb{R}^d$, write

$$[v]_{1:k} = \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix},$$

for the first k co-ordinates of v, where $k \leq d$. For $A \in \mathbb{R}^{d\times d}$, write $A^{(k)}$ for the leading principal minor of order k (i.e. the top left square submatrix of size $k \times k$).

Then,

$$\sqrt{n}\left([\hat{\theta}^{\mathrm{ML}}]_{1:k} - [\theta]_{1:k}\right) \xrightarrow{d} \mathcal{N}(0, I_\theta^{-1(k)}).$$

Consider testing $H_0 : g([\theta]_{1:k}) = 0$. One issue with applying the previous calculations is that $I_\theta^{-1(k)}$ depends on all of the co-ordinates of θ, so it is not necessarily constant if we fix the first k co-ordinates. The fix is to take $I_{\hat{\theta}^{\mathrm{ML}}}^{-1\ (k)}$ and use the following linear algebra fact: If

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

is symmetric, positive definite and $M = A^{-1}$ then $M_{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}$. This fact allows us to prove consistency of $\left[I_{\hat{\theta}^{\mathrm{ML}}}^{-1\ (k)}\right]$.

16.4 The Rao/Score Test

Instead of using the MLE, we could use the score function. We know that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \log p_\theta(X_i) \xrightarrow[d]{p_\theta^{\otimes n}} \mathcal{N}(0, I_\theta),$$

and

$$\frac{1}{n}\left(\sum_{i=1}^{n} \nabla_\theta \log p_\theta(X_i)\right)^{\mathsf{T}} I_\theta^{-1} \left(\sum_{i=1}^{n} \nabla_\theta \log p_\theta(X_i)\right) \xrightarrow[p_\theta^{\otimes n}]{d} \chi_d^2,$$

where d is the dimension of θ. We can establish consistency and asymptotic local power with a similar derivation to the Wald test.

16.4.1 Nuisance parameters (in modern research)

Nuisance parameters are common in high dimensional and causal inferences, amongst other areas. See [CCD+18] for general strategies on dealing with nuisance parameters.

add-on One approach to deal with nuisance parameters is to replace them with a consistent (under the null) estimator (e.g. their MLE) in the test statistic.

17 Lecture 25/3

17.1 Generalised Likelihood Ratio Tests

In Stat211, we say that for a simple null $H_0 : \theta = \theta_0$ versus a simple alternative $H_1 : \theta = \theta_1$, we can construct the likelihood ratio

$$R_n(\theta_1, \theta_0) = \frac{L_n(\theta_1)}{L_n(\theta_0)},$$

and the Neyman-Pearson lemma states that the test which rejects the null for large values of $L_n(\theta_1, \theta_0)$ is universally most powerful (UMP). The generalised likelihood ratio test (GLRT) extends this idea to composite null versus composite alternative testing.

Definition 17.1. Let $X \sim P_\theta$, for $\theta \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_1$. We want to test

$$H_0 : \theta \in \Theta_0 \text{ vs } H_1 : \theta \in \Theta_1.$$

The GLRT has test statistic

$$R_n = \frac{\sup_{\theta \in \Theta} p_\theta(x)}{\sup_{\theta \in \Theta_0} p_\theta(x)},$$

and critical function

$$\phi_{\text{GLRT}}(x) = \begin{cases} 1 & \text{if } R_n > k, \\ 0 & \text{if } R_n < k, \\ \gamma & \text{if } R_n = k. \end{cases}$$

k and γ are chosen so that the level constraint of the test is satisfied.

The numerator of R_n is $p_{\hat{\theta}^{\mathrm{ML}}}(x)$ and the denominator is the maximum likelihood when restricting to the null hypothesis. So intuitively, if the fraction is large then it is much more likely that the data is drawn from the alternative distribution than the null.

Example 17.2.

1. Consider $X = (X_1, \ldots, X_p)^\mathsf{T} \sim \mathcal{N}(\theta, \Sigma)$ where $\theta \in \mathbb{R}^p$ and Σ is known. The goal is to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The test statistic is

$$R_n = \frac{\sup_{\theta \in \mathbb{R}^p} \exp\left[-\frac{1}{2}(X-\theta)^\mathsf{T}\Sigma^{-1}(X-\theta)\right]}{\exp\left[-\frac{1}{2}X^\mathsf{T}\Sigma^{-1}X\right]}.$$

   The numerator is maximised when $X = \theta$, so $R_n = \exp\left(\frac{1}{2}X^\mathsf{T}\Sigma^{-1}X\right)$. Hence, we can write the critical region as $\{x : x^\mathsf{T}\Sigma^{-1}x > k'\}$. But what should k' be? We need that

$$\mathbb{P}_{H_0}\left[X^\mathsf{T}\Sigma^{-1}X > k'\right] = \alpha.$$

   We know that under H_0, $X^\mathsf{T}\Sigma^{-1}X \sim \chi^2_p$. So set k' as the $1-\alpha$ quantile of the χ^2_p distribution.

2. Goodness of fit testing: Consider $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Bern}(p)$ and test $H_0 : p = \pi$ versus $H_1 : p \neq \pi$. The test statistic is

$$R_n = \frac{\sup_{p \in [0,1]} p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}}{\pi^{\sum_{i=1}^n X_i}(1-\pi)^{n-\sum_{i=1}^n X_i}}.$$

   The numerator is maximised at the MLE $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$, so

$$R_n = \frac{\hat{p}^{n\hat{p}}(1-\hat{p})^{n(1-\hat{p})}}{\pi^{n\hat{p}}(1-\pi)^{n(1-\hat{p})}}.$$

   Two equivalent test statistics are

$$R_n' = n\hat{p}\left[\log\hat{p} - \log\pi\right] + n(1-\hat{p})\left[\log(1-\hat{p}) - \log(1-\pi)\right],$$

   and

$$R_n'' = \hat{p}\log\frac{\hat{p}}{\pi} + (1-\hat{p})\log\frac{1-\hat{p}}{1-\pi}.$$

This is the KL divergence between Bern($\hat{p}$) – the empirical distribution of the data – and Bern(π)! This is to be expected, since the MLE can be interpreted as a KL projection.

How should we set the cut off for R''_n? We need to derive the null distribution of R''_n. This is difficult. But we can understand the asymptotics of the GLRT in a general class of families. Hence we can get cutoffs (for large sample sizes) for the GLRT statistic without working out the null distribution on a case-by-case basis.

There are two basic questions about the GLRT which we should answer: Is the GLRT consistent? And what is its asymptotic local power?

17.1.1 Consistency of the GLRT

To resolve the question of consistency of the GLRT, we need to understand how to set the cutoff to achieve the right significance level. To do this, we need the asymptotic null distribution of the GLRT.

Theorem 17.3 (12.4.2 of [TSH] – Wilk's theorem)**.** *Let $X_1, \ldots X_n \overset{iid}{\sim} p_\theta$ where $\{p_\theta : \theta \in \Theta\}$ with partition $\Theta = \Theta_0 \cup \Theta_1$. Assume that the family is QMD; Θ is an open subset of $\mathbb{R}^k$ and I_θ is positive definite.*

(i) Under a simple-vs-simple testing problem: – $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ – we have

$$2 \log R_n \xrightarrow[H_0]{d} \chi^2_k.$$

(ii) Suppose Θ_0 can be represented as

$$\Theta_0 : \left\{ \theta : (g_1(\theta), \ldots, g_p(\theta))^\mathsf{T} = 0 \right\},$$

for some continuously differentiable $g_i : \mathbb{R}^k \to \mathbb{R}$ (i.e. Θ_0 is the nullset of $\boldsymbol{g}$). Define $D = D(\theta) \in \mathbb{R}^{p \times k}$ by

$$D_{ij} = \frac{\partial g_i}{\partial \theta_j}.$$

If $\operatorname{rank} D = p$ *then*

$$2 \log R_n \xrightarrow[H_0]{d} \chi^2_p.$$

This Theorem directly answers how to set the significant level. You can also use this to prove consistency of the GLRT (see Assignment 4, Question 3).

Proof of (i). We know that

$$2 \log R_n = 2 \sum_{i=1}^{n} \left[\log p_{\hat{\theta}^{\mathrm{ML}}}(X_i) - \log p_{\theta_0}(X_i) \right].$$

The regularity conditions imply that $\hat{\theta}^{\mathrm{ML}}$ is consistent for θ_0 under $H_0 : \theta = \theta_0$.

For this proof, we assume further smoothness conditions: $\theta \mapsto p_\theta$ is thrice continuously differentiable. Then Taylor expand around $\hat{\theta}^{\mathrm{ML}} = \theta_0$ (assuming the parameter is univariate):

$$2 \log R_n = 2 \sum_{i=1}^{n} \left[\frac{\partial}{\partial \log p_\theta} \bigg|_{\theta=\theta_0} \left(\hat{\theta}^{\mathrm{ML}} - \theta_0 \right) + \frac{\partial^2}{\partial \theta^2} \log p_\theta \bigg|_{\theta=\theta_0} \frac{\left(\hat{\theta}^{\mathrm{ML}} - \theta_0 \right)^2}{2} \right] + o_{p_{\theta_0}^{\otimes n}}(1). \tag{47}$$

We know that

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta \bigg|_{\theta = \hat{\theta}^{\mathrm{ML}}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta \bigg|_{\theta=\theta_0} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p_\theta \bigg|_{\theta=\theta_0} \left(\hat{\theta}^{\mathrm{ML}} - \theta_0 \right) + p_{p_{\theta_0}^{\otimes n}}(1), \end{aligned}$$

by the first order Taylor expansion of $\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log p_\theta$ around $\theta = \theta_0$.

Hence

$$\sqrt{n} \left(\hat{\theta}^{\mathrm{ML}} - \theta_0 \right) = \frac{-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta \big|_{\theta=\theta_0} + o_{p_{\theta_0}^{\otimes n}}(1)}{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p_\theta \big|_{\theta=\theta_0}}. \tag{48}$$

Plug (48) into (47):

$$2 \log R_n = \frac{-2 \left[\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta \big|_{\theta\theta_0} \right]^2}{\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p_\theta \big|_{\theta=\theta_0}} + \frac{\left[\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta \big|_{\theta=\theta_0} \right]^2}{\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p_\theta \big|_{\theta=\theta_0}} + o_{p_{\theta_0}^{\otimes n}}(1)$$

$$= \frac{-\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log p_\theta\big|_{\theta=\theta_0}\right)^2}{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log p_\theta\big|_{\theta=\theta_0}} + o_{p_{\theta_0}^{\otimes n}}(1)$$
$$\xrightarrow[H_0]{d} \left(\sqrt{I_{\theta_0}^{-1}}\mathcal{N}(0, I_{\theta_0})\right)^2$$
$$\sim \chi_1^2. \qquad \square$$

There are three properties that are crucial for the above proof:

1) The consistency of the MLE is necessary so that the Taylor approximation term is $o_{p_{\theta_0}^{\otimes n}}(1)$;
2) The normalised score is asymptotically Gaussian;
3) The negative gradient converges in probability to the Fisher information.

In setups where these properties hold, $2\log R_n$ will converge to a χ^2 random variable.

17.1.2 The asymptotic local power of the GLRT

To ascertain the asymptotic local power of the GLRT – assuming a simple null – we need to understand the distribution of $2\log R_n$ under $p_{\theta_0+h/\sqrt{n}}^{\otimes n}$. By Theorem 11.1, it suffices to understand the weak convergence of

$$\left(2\log R_n, \log\frac{dp_{\theta_0+h/\sqrt{n}}^{\otimes n}}{dp_{\theta_0}^{\otimes n}}\right), \tag{49}$$

under $p_{\theta_0}^{\otimes n}$. Yet

$$2\log R_n = 2\sum_{i=1}^n \log\frac{p_{\hat\theta^{\mathrm{ML}}}(X_i)}{p_{\theta_0}(X_i)} = 2\log\frac{p_{\hat\theta^{\mathrm{ML}}}^{\otimes n}}{p_{\theta_0}^{\otimes n}}.$$

So we need to look at the log-likelihood at $\hat\theta^{\mathrm{ML}}, \theta_0$ and $\theta_0 + h/\sqrt{n}$.

Assume that the model family is QMD at θ_0; the parameter space is an open subset of $\mathbb{R}^k$ and that I_{θ_0} is positive definite. The natural approach is to recall all of the expansions of $\log R_n$ that we have just computed. Then we can compute the joint distribution of (49). After some calculations, we arrive at

$$2\log R_n \xrightarrow[p_{\theta_0+h/\sqrt{n}}^{\otimes n}]{d} X^\mathsf{T} I_{\theta_0} X,$$

where $X \sim \mathcal{N}(h, I_{\theta_0}^{-1})$.add-on This implies $2 \log R_n \xrightarrow[p_{\theta_0+h/\sqrt{n}}^{\otimes n}]{d} \chi_k^2\left(\|h\|_2^2\right)$, where k is dimension of θ.

This connects with earlier results on estimators: Before we saw (Theorem 13.2) that the local asymptotic distribution of an estimator can be written in terms of a function of a single Gaussian $X \sim \mathcal{N}(h, I_{\theta_0}^{-1})$. Now we observe a similar phenomenon: the local asymptotic distribution of the log-likelihood ratio is the log-likelihood ratio of a single $X \sim \mathcal{N}(h, I_{\theta_0}^{-1})$.

In summary, the log-likelihood ratio before under the local alternative in QMD families effectively reduces to studying a corresponding testing problem in the Gaussian location model.

17.1.3 An example where the GLRT isn't nice

Consider a high-dimensional regression scenario where

1. $X(n) \in \mathbb{R}^{n \times p(n)}$,
2. $y(n) \in \mathbb{R}^{n \times 1}$,
3. $\beta^\star(n) \in \mathbb{R}^{p(n)}$, where $p(n) \to \infty$ as $n \to \infty$.

Suppose that $y(n) = f(X(n)\beta^\star(n), U)$, where U is some independent uniform. For example

$$y_i(n) = \mathbb{1}\{U \leq \exp\left[X_i(n)\beta^\star(n)\right]\}.$$

(This setup covers all the logistic regression examples.) Suppose that

$$\left[\beta^\star(n)\right]_{1:(n-1)} = \beta^\star(n-1),$$

where $[x]_{1:m}$ denotes the vector of the first m co-ordinates of x. We want to test

$$H_0 : \left[\beta^\star(n)\right]_1 = 0 \text{ versus } H_1 : \left[\beta^\star(n)\right]_1 \neq 0.$$

The likelihood ratio is given by

$$R_n = \frac{\sup_{\beta \in \mathbb{R}^{p(n)}} L(\beta | X(n), y(n))}{\sup_{\substack{\beta \in \mathbb{R}^{p(n)} \\ [\beta(n)]_1 = 0}} L(\beta | X(n), y(n))}.$$

We have $p(n)-1$ nuisance parameters: $[\beta^\star(n)]_{2:p(n)}$. The number of nuisance parameters is diverging to infinity.

An issue in this set-up is that the MLE is inconsistent. So we can't apply the classical derivations as above. We can however fall back on leave-one-out analysis and the asymptotic influence function, under certain relationships between $p(n)$ and n.

Suppose also that $\frac{p(n)}{n} \to k$ with $0 \leq k \leq 1$. The $\log R_n = l(\hat{\beta}) - l(\tilde{\beta})$, where $\hat{\beta}$ is the MLE of $\beta^\star$ in the full model and $\tilde{\beta}$ is the MLE in the reduced model without β_1. Using the facts that

$$\nabla l_{n,p}(\hat{\beta}) = \nabla l_{n,p-1}(\tilde{\beta}) = 0,$$

we can simplify

$$\log R_n = \frac{\hat{\beta}_1^2}{\frac{1}{p}\mathrm{Tr}\left(\left[\nabla^2 l(\tilde{\beta})\right]^{-1}\right)} + o_{p_{\theta_0}^{\otimes n}}(1).$$

$\nabla^2 l(\tilde{\beta})$ is the Hessian of the full log-likelihood evaluated at the reduced MLE. It looks reminiscent of the Fisher information.

Under $H_0 : \beta_1^\star = 0$, one can show

$$\sqrt{p}\hat{\beta}_1 \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

and

$$\mathrm{Tr}\left(\left[\nabla^2 l(\tilde{\beta})\right]^{-1}\right) \xrightarrow{P} \lambda,$$

where λ is a constant. Hence

$$\log R_n \xrightarrow{d} \frac{\sigma^2}{\lambda}\chi_1^2.$$

So we have almost recovered the asymptotic null distribution of the GLRT, modulo a rescaling factor.

See [SC18] for more details; generalisations beyond logistic regression are in a forthcoming publication.

18 Lecture 30/3

Recall our progress on understanding the asymptotic properties of hypothesis testing: so far, we have covered

1. consistency of tests;

2. the limiting local power (and its motivation);

3. asymptotic relative efficiency (ARE);

4. the Wald, score and GLR tests.

Today we want to understand how to pin down which test is optimal in a broad family of tests, or in a certain given parametric situation.

18.1 Asymptotic optimality of hypothesis tests

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$ and the one-sided test:

$$H_0 : \theta = \theta_0 \text{ versus } \theta > \theta_0.$$

In Stat211, we used 'uniformly most powerful' (UMP) as an optimality criterion in this setting. But in many scenarios the UMP test does not exist. As usual, we go to asymptotics to build a more general theory. We will show that with weak smoothness assumptions, asymptotically optimal tests do exist.

To characterise tests, we know we should use local power, since local alternatives allow for situations where there is non-trivial limiting power. So asymptotic optimality is most naturally studied in local settings.

18.1.1 Simple-vs-simple testing

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_{n,h} = \theta_0 + h/\sqrt{n}$. Under the Neyman-Pearson paradigm, for a sequence $\{\alpha_n\}$ of levels converging to α, the NP lemma says that the MP test $\phi_{n,h}$ at level α_n rejects when the likelihood ratio

$$L_{n,h} = \frac{p^{\otimes n}_{\theta_0+h/\sqrt{n}}}{p^{\otimes n}_{\theta_0}},$$

is sufficiently large. Specifically,

$$\phi_{n,h} = \begin{cases} 1 & \text{if } \log L_{n,h} > c_{n,h}, \\ \gamma_{n,h} & \text{if } \log L_{n,h} = c_{n,h}, \\ 0 & \text{if } \log L_{n,h} < c_{n,h}, \end{cases} \tag{50}$$

where the constants $c_{n,h}, \gamma_{n,h}$ are determined so that $\mathbb{E}_{\theta_0}\phi_{n,h} = \alpha_n$.

Lemma 18.1 (13.3.1 of [TSH]). *Assume a simple-vs-simple testing problem where $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_0 + h/\sqrt{n}$ with levels $\{\alpha_n\}$ converging to $\alpha \in (0,1)$. Suppose that $\{p_\theta : \theta \in \Theta\}$ is QMD at θ_0, with Ω an open subset of $\mathbb{R}$.*

1. *As $n \to \infty$,*
$$c_{n,h} \to \frac{-h^2 I_{\theta_0}}{2} + h I_{\theta_0}^{1/2} z_{1-\alpha}.$$
2. *$\mathbb{P}_{\theta_0}[\log L_{n,h} > c_{n,h}] \to \alpha$ so that $\mathbb{P}_{\theta_0}[\log L_{n,h} = c_{n,h}] \to 0$.*
3. *The power of $\phi_{n,h}$ satisfies*
$$\mathbb{E}_{p^{\otimes n}_{\theta_0 + h/\sqrt{n}}}\phi_{n,h} \to 1 - \Phi\left[z_{1-\alpha} - h I_{\theta_0}^{1/2}\right].$$

3. can be generalised to the scenario with alternatives $H_1 : \theta = \theta_0 + h_n/\sqrt{n}$ where $h_n \to h$ as long as $|h| < \infty$.

18.1.2 Asymptotically most powerful (AMP) tests

Definition 18.2. For testing $H_0 : \theta = \theta_0$ versus $\theta = \theta_n$, $\{\phi_n\}$ is *asymptotically most powerful* (AMP) at asymptotic level α if

(i) $\limsup_{n\to\infty} \mathbb{E}_{\theta_0}(\phi_n) \le \alpha$; and

(ii) For any other sequence of test functions $\{\psi_n\}$ satisfying (i),
$$\limsup_{n\to\infty} \mathbb{E}_{\theta_n}[\psi_n - \phi_n] \le 0.$$

I find it more intuitive to write property (ii) as

$$\liminf_{n\to\infty} \mathbb{E}_{\theta_n}\left[\phi_n - \psi_n\right] \geq 0.$$

Theorem 18.3 (13.3.1 of [TSH]). *Suppose $\{p_\theta : \theta \in \Theta\}$ is QMD at θ_0 with Θ an open subset of $\mathbb{R}$. Given $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$, test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_0 + h_n/\sqrt{n}$ where $h_n \to h > 0$. Then $\phi_n = \phi_n(X_1, \ldots, X_n)$ is AMP at level α if and only if $\mathbb{E}_{\theta_0}\phi_n \to \alpha$ and*

$$\limsup_{n\to\infty} \mathbb{E}_{\theta_0 + h_n/\sqrt{n}}\phi_n = 1 - \Phi\left(z_{1-\alpha} - hI_{\theta_0}^{1/2}\right).$$

This is what we would expect since this is the limiting power of the Neyman-Pearson MP test.

Unfortunately, the AMP $\phi_{n,h}$ (from (50) can depend on the value of h! With different choices of h, we are looking at different parts of the local neighbourhood around θ_0. The choice of h is usually arbitrary. So we would like a sequence $\{\phi_n\}$ that doesn't depend on h. Ideally, we want a sequence $\{\phi_n\}$ test which is AMP for "all possible choices of h". (We will make this precise later.) The test (50) based on the log-likelihood ratio does not necessarily satisfy this.

18.1.3 The score test

In QMD families, the (local) log-likelihood ratio satisfies

$$\log L_{n,h} = \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\theta_0}(X_i) - \frac{1}{2}h^2 I_{\theta_0} + o_{p_{\theta_0}^{\otimes n}}(1).$$

The log-likelihood ratio depends on the data X_i only through the QMD η_{θ_0}. And η doesn't depend on h. So a test based on η is an obvious candidate for a test which is AMP "for all possible h".

Definition 18.4. Define the *score statistic*,

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\theta_0}(X_i).$$

and the *score test*,

$$\tilde{\phi}_n = \begin{cases} 1 & \text{if } Z_n > I_{\theta_0}^{1/2} z_{1-\alpha}, \\ 0 & \text{otherwise.} \end{cases}$$

Why do we expect $\tilde{\phi}_n$ to be AMP "for all possible h"? We know that if $X_n - Y_n = o_{P_n}(1)$ and $Q_n \triangleleft P_n$ then $X_n - Y_n = o_{Q_n}(1)$ [TSH, Prob. 2.24]. Hence

$$\log L_{n,h} = \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\theta_0}(X_i) - \frac{1}{2} h^2 I_{\theta_0} + o_{p^{\otimes n}_{\theta_0 + h/\sqrt{n}}}(1),$$

in QMD families. (This follows since we proved contiguity of $p^{\otimes n}_{\theta_0+h/\sqrt{n}}$ in QMD families.) Hence the power of the LRT $\phi_{n,h}$ is derived from $\eta_{\theta_0}(X_i)$. Since $\phi_{n,h}$ is AMP, we would expect $\tilde{\phi}_n$ to be as well. Yet $\tilde{\phi}_n$ is not dependent on h, so its power corresponds to the power of $\phi_{n,h}$ for every h.

Lemma 18.5 (13.3.2 of [TSH])**.** *Suppose* $\{p_\theta : \theta \in \Theta\}$ *is QMD at* θ_0 *where* Θ *is an open subset of* $\mathbb{R}$*. Consider testing* $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta = \theta_0 + h/\sqrt{n}$ *with significance levels* $\alpha_n \to \alpha$*. Then the score test* $\tilde{\phi}_n$ *is asymptotically level* α *and for any* $0 < c < \infty$*,*

$$\sup_{0 \le h \le c} \left| \mathbb{E}_{\theta_0 + h/\sqrt{n}} \tilde{\phi}_n - \left[1 - \Phi\left(z_{1-\alpha} - h I_{\theta_0}^{1/2} \right) \right] \right| \xrightarrow{n \to \infty} 0.$$

So the limiting power of the score test against the local alternatives converges to the optimal limiting power uniformly in $h \in [0, c]$ for any $c > 0$. But the Lemma does not imply that the score test is universally optimal across all h, with h unconstrained in $\mathbb{R}$.

18.1.4 Asymptotically uniformly most powerful (AUMP)

Definition 18.6. Consider the one-sided simple-vs-composite testing: $H_0 : \theta = \theta_0$ versus $\theta > \theta_0$. A sequence of tests $\{\phi_n\}$ is *asymptotically uniformly most powerful* (AUMP) at (asymptotic) level α if

(i) $\limsup_{n\to\infty} \mathbb{E}_{\theta_0} \phi_n \le \alpha$, and

(ii) For any other sequence of tests $\{\psi_n\}$ satisfying (i),

$$\limsup_{n\to\infty} \sup_{\theta>\theta_0} \mathbb{E}_\theta \left[\psi_n - \phi_n\right] \leq 0.$$

So ϕ_n is AUMP if and only if it is asymptotically level α and it is AMP against any sequence of alternatives $\{\theta_n\}$ with $\theta_n > \theta$.

We can reparametrise property (ii) to be the supremum over local alternatives:

$$\limsup_{n\to\infty} \sup_{h>0} \mathbb{E}_{\theta_0+h/\sqrt{n}} \left[\psi_n - \phi_n\right] \leq 0.$$

So this is the natural extension of AMP to simple-vs-composite tests.

AUMP is typically too strong a definition. In most cases, an AUMP test will not exist. We can weaken AUMP by replacing the supremum over $h > 0$ with a supremum over compact sets (as in Lemma 18.5).

18.1.5 Locally asymptotically uniformly most powerful (LAUMP)

Definition 18.7. Consider the same setup as in Definition 18.6. A sequence of test $\{\phi_n\}$ is *locally asymptotically uniformly most powerful* (LAUMP) at (asymptotic) level α if

(i) $\limsup_{n\to\infty} \mathbb{E}_{\theta_0}\phi_n \leq \alpha$, and

(ii) For any other sequence of tests $\{\psi_n\}$ satisfying (i) and any $0 < c < \infty$,

$$\limsup_{n\to\infty} \sup_{0<\theta\leq\theta_0+c/\sqrt{n}} \mathbb{E}_\theta \left[\psi_n - \phi_n\right] \leq 0.$$

This is exactly the definition of AUMP except $\sup_{\theta>\theta_0}$ is changed to $\sup_{0<\theta\leq\theta_0+c/\sqrt{n}}$ for all $0 < c < \infty$.

As in the definition of AUMP, property (ii) can be reparametrised as

$$\limsup_{n\to\infty} \sup_{0<h\leq c} \mathbb{E}_{\theta_0+h/\sqrt{n}} \left[\psi_n - \phi_n\right] \leq 0.$$

18.1.6 Asymptotic optimality results in simple-vs-composite testing

Theorem 18.8 (13.3.2 of [TSH]). *Consider testing $H_0 : \theta = \theta_0$ verus $H_1 : \theta > \theta_0$ in a family which is QMD at θ_0 with non-zero Fisher information I_{θ_0} at θ_0. If $\phi_n = \phi_n(X_1, \dots, X_n)$ is a sequence of tests such that $\mathbb{E}_{\theta_0}\phi_n \to \alpha$ then*

(i) $\limsup_{n\to\infty} E_{\theta_0+h/\sqrt{n}} \leq 1 - \Phi\left(z_{1-\alpha} - hI_{\theta_0}^{1/2}\right)$*, for any h. (This follows by working with the Neyman-Pearson MP test from earlier.)*

(ii) ϕ_n is AUMP at level α if and only if

$$\sup_{h>0}\left|\mathbb{E}_{\theta_0+h/\sqrt{n}}\phi_n - \left[1 - \Phi\left(z_{1-\alpha} - hI_{\theta_0}^{1/2}\right)\right]\right| \to 0.$$

(iii) ϕ_n is LAUMP at level α if and only if, for all $c > 0$,

$$\sup_{0<h\leq c}\left|\mathbb{E}_{\theta_0+h/\sqrt{n}}\phi_n - \left[1 - \Phi\left(z_{1-\alpha} - hI_{\theta_0}^{1/2}\right)\right]\right| \to 0.$$

The score test $\tilde{\phi}_n$ is LAUMP in QMD families (from earlier calculations). The following Theorem states that we can use the score test as a reference.

Theorem 18.9 (13.3.3 of [TSH]). *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ in a family which is QMD at θ_0 with derivative η_{θ_0} and non-zero Fisher information I_{θ_0} at θ_0. Define the score test*

$$\tilde{\phi}_n = \begin{cases} 1 & \text{if } Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_{\theta_0}(X_i) \geq I_{\theta_0}^{1/2} z_{1-\alpha}, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

(i) $\tilde{\phi}_n$ is LAUMP at level α;

(ii) Sufficient condition: Any test sequence ψ_n satisfying $\psi_n - \tilde{\phi}_n \xrightarrow{P} 0$ under p_{θ_0} is also LAUMP at level α.

(iii) Necessary condition: If a test sequence $\{\psi_n\}$ is LAUMP at level α, then $\psi_n - \tilde{\phi}_n \xrightarrow{P} 0$ under p_{θ_0}.

(iv) If, in addition, $Z_n \to \infty$ in $p_{\theta_n}^{\otimes n}$-probability whenever

$$\sqrt{n}\,(\theta_n - \theta_0) \to \infty,$$

then the score $\tilde{\phi}_n$ is also AUMP at level α.

Note that (ii) does not require anything of the behaviour of ψ_n under the alternative hypothesis! Why can (ii) be true, despite this? Since we are looking at LAUMP, we only need to consider a local neighbourhood of θ_0. Further, the behaviour in this local neighbourhood is governed by p_{θ_0} since the family is QMD at θ_0.

Statement (iv) is saying that if you take a sequence $\{\theta_n\}$ which goes to θ_0 slower than rate $n^{-1/2}$ and

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\theta_0}(X_i) \to \infty,$$

when $X_i \overset{iid}{\sim} p_{\theta_n}$, then the score test $\tilde{\phi}_n$ is AUMP.

What does $Z_n \to \infty$ in $p_{\theta_n}^{\otimes n}$-probability mean? It means that for all $C \in \mathbb{R}$,

$$\mathbb{P}_{p_{\theta_n}^{\otimes n}}\left(Z_n < C\right) \to 0.$$

19 Lecture 1/4

19.1 Examples of asymptotic optimality

Example 19.1. Consider the Laplace density

$$p_\theta(x) = \frac{1}{2}\exp\left(-|x-\theta|\right),$$

and test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. The QMD is given by

$$\eta_\theta(x) = \begin{cases} \frac{p'_\theta(x)}{p_\theta(x)} & \text{if } p_\theta(x) > 0 \text{ and } p'_\theta(x) \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$

from Section 3. Under the null,

$$\eta_0(X_i) = \text{sign}(X_i) \text{ and } I_0 = 1,$$

so the score statistic is

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \operatorname{sign}(X_i),$$

and we reject if $Z_n > z_{1-\alpha}$.

The score test is LAUMP by Theorem 18.9. To determine if it is AUMP, use Theorem 18.9 again: We need to check whether $Z_n \to \infty$ in $p_{\theta_n}^{\otimes n}$-probability whenever $\sqrt{n}\theta_n \to \infty$. We can compute

$$\operatorname{Var}_{\theta_n} Z_n = \operatorname{Var} \operatorname{sign} X_i \leq \mathbb{E}\left[\operatorname{sign} X_i\right]^2 = 1,$$

and

$$\begin{aligned}
\mathbb{E}_{\theta_n} Z_n &= \sqrt{n}\mathbb{E}_{\theta_n} \operatorname{sign}(X_i) \\
&= \sqrt{n}\left[\mathbb{P}_{\theta_n}\left(X_i > 0\right) - \mathbb{P}_{\theta_n}\left(X_i < 0\right)\right] \\
&= \sqrt{n}\left(1 - e^{-\theta_n}\right) \\
&\to \infty,
\end{aligned}$$

as $\sqrt{n}\theta_n \to \infty$ by L'Hôpital's rule. Chebychev's inequality then proves $Z_n \xrightarrow[p_{\theta_n}^{\otimes n}]{P} \infty$.

Example 19.2. Suppose

$$X_i = (U_i, V_i) \overset{iid}{\sim} \operatorname{MVN}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

and consider testing $H_0 : \rho = 0$ against $H_1 : \rho > 0$. The score statistic is given by

$$\begin{aligned}
Z_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \rho} \log L_n(\rho)\bigg|_{\rho=0} \\
&= \ldots = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i V_i.
\end{aligned}$$

We can use the same reasoning as in the previous example to show that the score test is AUMP:

$$\mathbb{E}_{\rho_n} Z_n = \sqrt{n}\rho_n,$$

and

$$
\begin{aligned}
\mathrm{Var}_{\rho_n}(Z_n) &= \mathrm{Var}_{\rho_n}\left(U_i V_i\right) \\
&\leq \mathbb{E}\left(U_i^2 V_i^2\right) \\
&= \mathbb{E}_{\rho_n}\left[V_i^2 \mathbb{E}_{\rho_n}\left(U_i^2 \middle| V_i^2\right)\right] \\
&= \mathbb{E}_{\rho_n}\left[V_i^2\left(\rho_n^2 V_i^2 + (1-\rho_n^2)\right)\right] \\
&= \rho_n^2 \mathbb{E} V_i^4 + (1-\rho_n^2)\mathbb{E} V_i^2 \\
&\leq 4,
\end{aligned}
$$

where the fourth line follows since $U_i | V_i \sim \mathcal{N}(\rho_n V_i, 1-\rho_n^2)$; and the last line follows since $0 \leq \rho_n \leq 1$ and $\mathbb{E}V_i^4 = 3$ and $\mathbb{E}V_i^2 = 1$.

19.2 Asymptotic optimality in two-sided testing

If we want to move to two-sided testing, then the notions of AUMP and LAUMP are typically unrealistically strong. (There will be no LAUMP test in most setups.) Instead, look at optimality within a smaller class of tests – tests which satisfy (local) asymptotic unbiasedness. See Homework 5 for more details.

19.3 Generalisations beyond QMD families

We have been studying one-sided testing in QMD families where the parameter space Θ is an open subset of $\mathbb{R}$. In practise, we will encounter problems with multivariate parameters, nuisance parameters and non-iid data. See [TSH] after Theorem 13.4.1 for examples.

We know that in simple-vs-simple hypothesis tests, the NP lemma shows that the log-likelihood ratio is UMP. When we wanted to look at the LAUMP, we realised that we needed a test that was uniform across any choice of h. This let to the score test in QMD families, since we could expand the local log-likelihood ratio in terms of the score. But we can study an equivalent test in a broader class of families.

Recall that local asymptotically normal families in section 12.2.1. Here we expanded the log-likelihood ratio in terms of some random vectors Δ_n. It turns out that we can replace the score statistic $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\theta_0}(X_i)$ with Δ_n and maintain many of the asymptotic optimality properties.

Definition 19.3. Let $\{Q_{n,h} : h \in \mathbb{R}^k\}$ be a family parametrised by $h \in \mathbb{R}^k$ where the subscript n denotes the density when n samples (not necessarily indepedent) are observed. Define

$$L_{n,h} = \frac{dQ_{n,h}}{dQ_{n,0}},$$

assuming that $Q_{n,h}$ has a density with respect to some base measure.

$\{Q_{n,h} : h \in \mathbb{R}^k\}$ is *asymptotically normal* if there exists a sequence of random vectors Z_n and covariance matrix C such that

$$\log L_{n,h} = h^\intercal Z_n - \frac{1}{2} h^\intercal C h + o_{Q_{n,0}}(1), \tag{51}$$

where

$$Z_n \xrightarrow[Q_{n,0}]{d} \mathcal{N}(0, C).$$

The RHS of (51) looks like the log-likelihood ratio $h^\intercal Z - \frac{1}{2} h^\intercal C h$ from observing $Z \sim \mathcal{N}(Ch, C)$.

Theorem 19.4 (13.4.1 of [TSH])**.** *Let $\{Q_{n,h} : h \in \mathbb{R}^k\}$ be an asymptotically Normal sequence of models with covariance matrix C and random vector Z_n. Let ϕ_n be a test – i.e. a function defined on the same probability space as $Q_{n,h}$, and taking values in $[0, 1]$. Let $\pi_n(h) = \mathbb{E}_{Q_{n,h}} \phi_n$ denote the power of ϕ_n against $Q_{n,h}$.*

For every subsequence $\{n_j\}$ of $\mathbb{N}$, there exists a further subsequence $\{n_{j_m}\}$ and a test ϕ based on $Z \sim \mathcal{N}(Ch, C)$ such that, for every h,

$$\pi_{n_{j_m}}(h) \to \pi(h),$$

as $m \to \infty$, where $\pi(h)$ is the power of ϕ.

Why is this Theorem useful? The idea is that you can correspond a test ϕ_n of $Q_{n,h}$ with a test ϕ of the limiting model $\mathcal{N}(Ch, C)$. So the UMP test in the limiting model $\mathcal{N}(Ch, C)$ gives an upper bound on the limiting power of tests ϕ_n. Further, if you can construct ϕ_n with limiting power equal to this upper bound, then ϕ_n is optimal in some sense. See Section 10 for details.

19.4 Minimax Testing

add-on This section was given as a guest lecture from Prof. Subhabrata Sen and is not assessable.

History: Minimax testing was initiated in the 1980's in the Soviet school (Ingster) and was developed in parallel to Le Cam's work.

Setup: Suppose a parametric model with $X_1, \dots X_n \overset{iid}{\sim} p_\theta$ where $\theta \in \mathbb{R}$. Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. We have studied the sequence of alternatives $\theta_n = \theta_0 + h_n$:

1. If $h_n = o(n^{-1/2})$ then "detection is impossible" – i.e. no test can do better than random guessing.
2. If $\sqrt{n}h_n \to \infty$ then "testing is easy" – i.e. there exists a test with both type 1 and type 2 errors going to zero as $n \to \infty$. (We call such a test *powerful*.)

So $h_n = h/\sqrt{n}$ is the right regime for these parametric problems – the testing problem is neither too easy nor too difficult. In this case, we can use all the theory we have developed earlier to answer this testing problem.

Question: what happens in non-parametric or high dimensional settings?

Example 19.5. The Gaussian sequence model: For $1 \leq i \leq n$¡ let $X_i = \theta_i + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. We want to test

$$H_0 : \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} = \mathbf{0}$$

against

$$H_1 : \boldsymbol{\theta} \in \Theta(s, A) = \{\tau \in \mathbb{R}^n : \tau \text{ is } s\text{-sparse and } \tau_i \geq A \text{ if } \tau_i \neq 0\},$$

where s-sparse means that at most s co-ordinates are non-zero. This is a high dimensional problem as $\boldsymbol{\theta}$ grows with n.

Example 19.6. Non-parametric regression: Suppose

$$y_I = f\left(\frac{i}{n}\right) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0,1)$ and $\frac{i}{n}$ is the fixed design values. Suppose $f \in \mathcal{H}(\alpha)$ (the Hölder class) and test

$$H_0 : f = 0 \text{ versus } H_1 : f \in \mathcal{H}(\alpha) \text{ such that } \|f\|_2^2 \geq \rho_n.$$

In these types of examples, we have to give up on the "fine properties" that we derived in parametric models and look at "coarse properties" of the testing problem. We need to think broadly about the separation parameters (in the above examples, these are ρ_n and (A, s)).

Following the approach taken in the parametric problem (that is, finding that the rate $h_n = h/\sqrt{n}$ allowed for non-trivial testing), can we identify the "minimum separation" required for non-trivial testing? (We will make this question precise.)

Definition 19.7. Given a test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_n$, the *risk* of a sequence of tests $\{T_n : n \in \mathbb{N}\}$ is defined as

$$\mathrm{Risk}(T_n, \Theta_n) = \mathbb{E}_{\theta_0} T_n + \sup_{\theta \in \Theta_n} \left(1 - E_\theta T_n\right).$$

The *minimax risk* of the testing problem is defined as

$$R_n(\Theta_n) = \min_{T_n} \mathrm{Risk}(T_n, \Theta_n).$$

The first term of the risk is the type 1 error rate and the second term is the type 2 error rate.

Definition 19.8. A test sequence $\{T_n : n \in \mathbb{N}\}$ is *asymptotically poweful* if $\mathrm{Risk}(T_n, \Theta_n) \to 0$ as $n \to \infty$.

We say that "detection is possible" if there exists a sequence of asymptotically powerful tests. So "the minimum separation for non-trivial testing" means "the minimum separation such that detection is possible".

If you can find a sequence of asymptotically powerful tests at a certain separation d, then d is an upper bound on the minimum separation. Actually determining the minimum separation is much harder.

20 Lecture 6/4

20.1 Bayesian Asymptotics

The reference for this section is [GR].

20.1.1 Framework and set-up

The hypothesis model is $\mathcal{M} = \{p_{\tilde{\theta}} : \tilde{\theta} \in \mathcal{H}\}$. The data generating process goes in two steps: First generate $\Theta \sim \pi(\cdot)$ (π is called the prior distribution) to get the realised parameter θ. Second, draw $Y_1, \ldots, Y_n \overset{iid}{\sim} p_\theta(\cdot)$. So there are random variables $(\Theta, Y_1, \ldots, Y_n)$ with joint density:

$$(\theta, y_1, \ldots, y_n) \mapsto p(\theta, y_1, \ldots, y_n) = \pi(\theta) \prod_{i=1}^{n} p_\theta(y_i).$$

The goal is to infer about the unknown (realised) parameter θ, via the posterior distribution

$$\pi(\theta | y_1, \ldots, y_n) = \frac{\pi(\theta) \prod_{i=1}^n p_\theta(y_i)}{\int_{\mathcal{H}} \pi(\nu) \prod_{i=1}^n p_\nu(y_i) d\nu}.$$

(This posterior distribution is only defined if the integral in the denominator is finite.) Denote the likelihood $\mathcal{L}_n : \nu \mapsto \prod_{i=1}^n p_\nu(y_i)$. The denominator $\int_{\mathcal{H}} \pi(\nu) \prod_{i=1}^n p_\nu(y_i) d\nu$ is the marginal likelihood (aka the marginal density of $y_1, \ldots, y_n$).

In Bayesian inference, all we care about is the posterior distribution. So the questions that make sense here are properties of the posterior as the sample size goes to infinity. We will describe properties of the posterior that are analogous to properties that we've been studying in the frequentist setting.

Remark 20.1. The posterior density might still exist even when the prior is not a proper density (i.e. $\int_{\mathcal{H}} \pi(\nu) d\nu = \infty$). Priors with infinity mass are called *improper priors*. An example is the flat prior, where $\pi(\nu)$ is constant for all $\nu \in \mathcal{H}$. (This is an improper prior if $\mathcal{H}$ is not bounded.) In this setting, there is no joint distribution for $(\Theta, Y_1, \ldots, Y_n)$ but there is a posterior distribution $\Theta | y_1, \ldots, y_n$.

20.1.2 Intepretation of the posterior

The posterior distribution can be characterised as the minimiser of the quantity

$$-\int_{\mathcal{H}} \log\left[\prod_{i=1}^{n} p_\theta(y_i)\right] \nu(\theta)d\theta + \int_{\mathcal{H}} \log\left[\frac{\nu(\theta)}{\pi(\theta)}\right] \nu(\theta)d\theta, \tag{52}$$

over all probability distributions ν satisfying some regularity conditions that ensure the above quantity exists. (This is proven formally in Homework 5.)

The first term of (52) is minimised by the probability distribution ν that maximises the likelihood. This can be viewed as an analogue to the MLE.

The second term of (52) is the minimiser of the KL divergence $\mathrm{KL}(\nu\|\pi)$. Thus, the posterior can be interpreted as simultaneously trying to optimise the likelihood while staying "true" to the prior.

This elucidates the Bayesian perspective to inference: update the prior information based on the data likelihood.

Bayesian inference can also be interpreted in terms of "entropic inference" [Cat11].

20.1.3 Consistency

We focus on the behaviour of the posterior as $n \to \infty$ in the context of parametric and finite dimensional models. This material doesn't carry across to the general, non-parametric setting. (For details here, see [GR] and [Rou16].)

We would intuitively expect that the posterior distribution would converge to a point if we had access to an infinite amount of data. That is, we expect that as $n \to \infty$, the posterior distribution would concentrate around some value $\theta^\star$. This idea is formalised by the notion of consistency.

Definition 20.2 (1.3.1 of [GR]). Assume that $Y_1, \ldots, Y_n \overset{iid}{\sim} p^\star$. A sequence of posterior distributions $\{\pi(d\theta|Y_1, \ldots, Y_n)\}_{n=1}^{\infty}$ is *consistent* at $\theta^\star$ if, $p^\star$-a.s., for every neighbourhood*** U of $\theta^\star$,

$$\pi(U|Y_1, \ldots, Y_n) \xrightarrow{n\to\infty} 1. \tag{53}$$

***A set containing an open set around $\theta^\star$

The $\forall$ neighbourhoods quantifier is within the a.s. quantifier – that is (53) is stating:

$$\mathbb{P}_{p^\star}\left(\{\pi(U|Y_1,\dots,Y_n)\to 1 : U \text{ neighbourhood of } \theta^\star\}\right)=1.$$

In metric spaces (where the countable set $\{B_{1/n}(\theta^\star) : n\in\mathbb{N}\}$ forms a base for neighbourhoods of $\theta^\star$), consistency is equivalent to

$$\pi(d(\theta,\theta^\star)\le\epsilon|Y_1,\dots,Y_n)\xrightarrow[n\to\infty]{p^\star\text{-a.s.}}1,$$

for all $\epsilon>0$. (So in metric spaces, we can swap the order of the $\forall$ and a.s. quantifiers.)

If a.s. is replaced with an 'in probability' statement, then the property is called *weak consistency.*

The following consistency theorem (in the well-specified setting) is due to Doob:

Theorem 20.3 (1.3.2 of [GR], 10.10 of [vdV])**.** *Assume that the model is identifiable – that is, the map $\mathcal{H}\ni\theta\mapsto[p_\theta]$ (where $[p]$ is the equivalence class of a.e.-equal densities) is injective. Further, suppose that $\mathcal{H}\subset\mathbb{R}^d$, $\mathcal{Y}\subset R^p$ and π is a proper prior.*

Then there exists $\mathcal{H}_0\subset\mathcal{H}$ such that $\pi(\mathcal{H}_0)=1$ and such that the posterior is consistent at every $\theta\in\mathcal{H}_0$ whenever the data generating distribution is p_θ.

Proof. Define $M_n=\pi(A|Y_1,\dots,Y_n)$, where A is a measurable subset of $\mathcal{H}$. We will show that M_n is a martingale with respect to the σ-algebra $\mathcal{F}_n=\sigma(Y_1,\dots,Y_n)$:

$$\begin{aligned}\mathbb{E}\left[M_{n+1}|\mathcal{F}_n\right]&=\mathbb{E}Y_{n+1}\left[\pi(A|Y_1,\dots,Y_{n+1})\right]\\&=\mathbb{E}_{Y_{n+1}}\left[\mathbb{E}_\Theta\left(\mathbb{1}\{\Theta\in A\}\,|Y_1,\dots,Y_{n+1}\right)\right]\\&=\mathbb{E}_\Theta\left[\mathbb{1}\{\Theta\in A\}\,|Y_1,\dots,Y_n\right]\\&=\pi(A|Y_1,\dots,Y_n)\\&=M_n,\end{aligned}\tag{54}$$

where the notation $\mathbb{E}_X g(X,Y)$ denotes the expectation $\mathbb{E}\left[g(X,Y)|Y\right]$ is over the random X, conditioning on the other variables Y; and (I think) the third line follows by Fubini-Tonelli (I think this is where we need that π is proper).

Thus, the sequence of posteriors M_n is a martingale and bounded between zero and one. Lévy's upward theorem implies that

$$\pi(A|Y_1, \ldots, Y_n) \xrightarrow{\text{a.s.}} n \to \infty \pi(A|Y_1, Y_2, \ldots).$$

With some work, we can use this to show that there exist measurable $\mathcal{H}_0$ with $\pi(\mathcal{H}_0) = 1$ such that for any measurable set A, if $\theta \in A \cap \mathcal{H}_0$, then

$$\pi(A|Y_1, Y_2, \ldots) = 1,$$

$p_\theta^{\otimes\infty}$-a.s. (we haven't properly defined this notation yet, but intuitively it is simply $p_\theta^{\otimes n}$ in the limit as $n \to \infty$). □

Where does the proof break down in the misspecified setting? Assume that $Y_1, \ldots, Y_{n+1} \overset{iid}{\sim} p^\star$. The LHS of (54) is

$$\mathbb{E}_{Y_{n+1}}\left[\pi(A|Y_1, \ldots, Y_{n+1})\right] = \int_{\mathbb{R}^p} \frac{\int_A \prod_{i=1}^{n+1} p_\theta(Y_i)\pi(d\theta)}{p(Y_{1:n+1})} p^\star(Y_{n+1}) dY_{n+1},$$

and the RHS is

$$\frac{\int_A \prod_{i=1}^{n} p_\theta(Y_i)\pi(d\theta)}{p(Y_{1:n})},$$

where $p(Y_{1:n}) = \int_{\mathcal{H}} p_\nu(Y_{1:n})\pi(\nu)d\nu$ is the marginal distribution.

The LHS and RHS are not equal unless special relations hold between $p^\star(Y_{n+1})$ and $p(Y_{1:n+1})$. (add-on To be honest, I don't really understand this argument.) So $\pi(A|Y_1, \ldots, Y_n)$ is not necessarily a martingale.

Theorem 20.4 (Wald type consistency, 1.3.4 of [GR]). *Let $\mathcal{H}$ be compact, $\{p_\theta : \theta \in \mathcal{H}\}$ be distributions on $\mathcal{Y}$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} p^\star$. Assume that $\theta \mapsto p_\theta(y)$ is continuous for all $y \in \mathcal{Y}$; and for all $\theta \in \mathcal{H}$, $y \mapsto p_\theta(y)$ measurable. Further suppose*

$$\int_{\mathcal{Y}} \sup_{\theta \in \mathcal{H}} |\log p_\theta(y)| p^\star(dy) < \infty.$$

Then the MLE converges to $\theta^\star$, $p^\star$-a.s., where $\theta^\star$ is assumed to be the unique maximiser of

$$l^\star : \theta \mapsto \int_{\mathcal{Y}} \log p_\theta(y) p^\star(dy).$$

Further, if $\theta^\star$ is in the support of the prior, then the posterior is consistent at $\theta^\star$, in the sense that $p^\star$-a.s.,

$$\pi(U|Y_1,\dots,Y_n) \to 1,$$

as $n \to \infty$, for any neighbourhood U of $\theta^\star$.

The first half of the Theorem is a frequentist result, which we have proved already. The second half is a Bayesian consistency-type result. However, the second half does not prove consistency in the sense of Definition 20.2, since it is a statement under $p^\star$, not $p_{\theta^\star}$.

This theorem shows that both the MLE and the posterior converge to the same limiting value $\theta^\star$. It is an important example of Bayesian-frequentist reconciliation.

21 Lecture 8/4

21.1 Wald-type consistency proof

Proof of Theorem 20.4. We will prove the Theorem only for metric spaces. Fix a neighbourhood U of $\theta^\star$. Our goal is to show that

$$\pi(U|Y_1,\dots,Y_n) = \frac{\int_U \prod_{i=1}^n p_\theta(Y_i)\pi(d\theta)}{\int_{\mathcal{H}} \prod_{i=1}^n p_\theta(Y_i)\pi(d\theta)} \xrightarrow{p^\star\text{-a.s.}} 1,$$

as $n \to \infty$. (Note that this is only sufficient to prove the Theorem, when we are working in a metric space.)

Write $\mathcal{H} = U \cup K$ where $K = U^c \cap \mathcal{H}$. Define

$$R = \frac{\int_K \prod_{i=1}^n p_\theta(Y_i)\pi(d\theta)}{\int_U \prod_{i=1}^n p_\theta(Y_i)\pi(d\theta)},$$

and observe that

$$\pi(U|Y_1,\dots,Y_n) = \frac{1}{1+R}.$$

Thus, it suffices to show that R converges to zero $p^\star$-a.s. To show this, we will find a lower bound on the denominator of R, and an upper bound on the numerator.

From the assumptions, the strong ULLN holds:

$$\sup_{\theta \in \mathcal{H}} \left| \frac{1}{n} l_n(\theta) - l^\star(\theta) \right| \xrightarrow[n\to\infty]{p^\star\text{-a.s.}} 0.$$

Define $U_\delta \subset U$ by

$$U_\delta = \{\theta : d(\theta, \theta^\star) < \delta\} \cap U,$$

for any $\delta > 0$. Let $L_1 = \sup_{\theta \in K} l^\star(\theta)$ and $L_2 = \inf_{\theta \in U_\delta} l^\star(\theta)$. We know that $l^\star(\theta^\star) > L_1, L_2$ by assumption. Also, as $\delta \to 0$, $L_2 \to l^\star(\theta^\star)$ by continuity of $l^\star$. Hence, for small enough δ, we have $L_2 > L_1$, or equivalents, $A_2 > A_1$, where $A_i = L_i - l^\star(\theta^\star)$.

Choose $\epsilon > 0$ such that $A_1 + \epsilon < A_2 - \epsilon$. From the strong ULLN, there exists $N \in \mathbb{N}$, such that, for all $n \geq N$ and all $\theta \in \mathcal{H}$,

$$\left| \frac{1}{n} l_n(\theta) - l^\star(\theta) \right| < \epsilon.$$

Thus, for all $\theta \in K$, we have $\frac{1}{n} l_n(\theta) < L_1 + \epsilon$ while for all $\theta \in U_\delta$, we have $\frac{1}{n} l_n(\theta) > L_2 - \epsilon$. This gives

$$\int_K \prod_{i=1}^n p_\theta(Y_i) \pi(d\theta) < \exp\left(n(L_1 + \epsilon)\right) \pi(K)$$
$$\int_{U_\delta} \prod_{i=1}^n p_\theta(Y_i) \pi(d\theta) > \exp\left(n(L_2 - \epsilon)\right) \pi(U_\delta).$$

Further, $\exp\left(n(L_2 - \epsilon)\right) \pi(U_\delta) > 0$, since $\theta^\star$ is in the support of π. Putting this together,

$$\begin{aligned} R &\leq \frac{\exp\left(n(L_1 + \epsilon)\right) \pi(K)}{\exp\left(n(L_2 - \epsilon)\right) \pi(U_\delta)} \\ &= \left[\exp\left(L_1 + \epsilon - (L_2 - \epsilon)\right)\right]^n \frac{\pi(K)}{\pi(U_\delta)} \\ &\xrightarrow{n\to\infty} 0, \end{aligned}$$

since $L_1 + \epsilon < L_2 - \epsilon$. □

21.2 Robustness (to the prior)

"Any nice prior will lead to the same posterior, given enough data."

Theorem 21.1 (Theorem 1.3.1 of [GR]). *Let $\{p_\theta : \theta \in \mathcal{H}\}$ be a well specified model and observe $Y_1, \ldots, Y_n \overset{iid}{\sim} p_{\theta^\star}$ for some $\theta^\star \in \int \mathcal{H}$. Let π_1, π_2 be two (proper) priors, which are positive and continuous at $\theta^\star$. If both posteriors are consistent at $\theta^\star$, then*

$$\int_{\mathcal{H}} |\pi_1(\theta|Y_{1:n}) - \pi_2(\theta|Y_{1:n})| d\theta \xrightarrow[n\to\infty]{p_{\theta^\star}\text{-}a.s.} 0.$$

Informally, this Theorem means that if two researchers perform Bayesian inference with different, but well behaved, priors and the same well specified model, then their conclusions will be the same for large amounts of data. For the misspecified setting, see [Rou16].

21.3 The Bernstein-von Mises theorem

Last lecture we say the Bayesian analogue of the frequentist notion of consistency. Now we will see that analogue of asymptotic Gaussianity in the Bayesian context.

Let Θ_n be a random variable with distribution $\pi(\cdot|Y_{1:n})$. Let $\hat{\theta}_n$ be the MLE and $T_n = \sqrt{n}\left(\Theta_n - \hat{\theta}_n\right)$. Denote the distribution of T_n by π_n.

The Bernstein-von Mises (BvM) theorem states that, under appropriate regularity conditions,

$$\int_{\mathcal{H}} \left| \pi_n(t) - \frac{1}{\sqrt{(2\pi)^k |V^\star|}} \exp\left(-\frac{1}{2} t^\top V^{\star-1} t\right) \right| dt \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0, \tag{55}$$

where k is the dimension of $\mathcal{H}$, $V^\star$ is some positive definite matrix and $p^\star$-a.s. is the true data generating distribution.

The second term of the integrand is the density of $\mathcal{N}(\mathbf{0}, V^\star)$. So the BvM theorem states that π_n converges in TV distance to $\mathcal{N}(\mathbf{0}, V^\star)$, in $p^\star$ probability. Informally, this means the posterior distribution behaves asymptotically like $\mathcal{N}(\hat{\theta}_n, V^\star/n)$.

This theorem is useful since it directly leads to the formulation of the Bayesian information criteria (BIC) for model selection. See the section for details.

21.3.1 What is $V^\star$?

We will see that

$$V^\star = \left[\mathbb{E}_\star\left(-\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^\star}\right)\right]^{-1},$$

where the expectation is with respect to the true data generating distribution $p^\star$. In the well-specified setting, $\theta^\star$ is the true data generating parameter, so that $p^\star = p_{\theta^\star}$. In the misspecified setting, $\theta^\star$ is the limit of $\hat{\theta}_n$ (i.e. the KL projection of $p^\star$ onto $\mathcal{M} = \{p_\theta : \theta \in \mathcal{H}\}$).

So the asymptotic variance of Θ_n is the same as the MLE $\hat{\theta}_n$ in the well specified setting. But it is different to the asymptotic variance of $\hat{\theta}_n$, given by the sandwich formula, in the misspecified setting. This means that the credible intervals will agree with the confidence intervals in the well specified case. (Why? Both intervals will eventually be centred at $\hat{\theta}_n$ and their widths will be the same, since the variances are equal.) But this is not so in the misspecified case.

21.3.2 Intuition for BvM

Let $t = \sqrt{n}\left(\Theta_n - \hat{\theta}_n\right)$. Taylor expanding,

$$l_n(\hat{\theta}_n + t/\sqrt{n}) \approx l_n(\hat{\theta}_n) + \frac{t}{\sqrt{n}}\nabla_\theta l_n(\hat{\theta}_n) + \frac{1}{2n}t^\intercal \nabla_\theta^2 l_n(\hat{\theta}_n)t.$$

We know that the second term on the RHS is zero by the property of the MLE and that the third term

$$\frac{1}{n}\nabla_\theta^2 l_n(\hat{\theta}_n) \to -J^\star = -\mathbb{E}_\star\left[-\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^\star}\right].$$

Hence

$$l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n) \approx -\frac{1}{2}t^\intercal J^\star t.$$

The idea here is that if the prior density is flat (i.e $\pi(\hat{\theta}_n + t/\sqrt{n}) - \pi(\hat{\theta}_n) \approx 0$) in small neighbourhoods around the MLE, then the posterior density is approximately Gaussian with variance $V^\star = [J^\star]^{-1}$.

21.3.3 Proof of BvM

Proof of (55). Assume the univariate setting $\mathcal{H} = \mathbb{R}$. (The multivariate setting is an easy extension.) We will state the assumptions as we go along, and recap them at the end.

Assume the required conditions for consistency of the MLE: $\hat{\theta}_n p^\star\text{-a.s.}\theta^\star$.

Let $t = \sqrt{n}\left(\theta - \hat{\theta}_n\right)$. The density $\pi_n(\cdot)$ of the random variable $\sqrt{n}\left(\Theta_n - \hat{\theta}_n\right)$ is given by

$$\begin{aligned}\pi_n(t) &= \frac{\pi(\hat{\theta}_n + t/\sqrt{n}) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n})\right)}{\int_{\mathbb{R}} \pi(\hat{\theta}_n + u/\sqrt{n}) \exp\left(l_n(\hat{\theta}_n + u/\sqrt{n})\right) du} \\ &= \frac{\pi(\hat{\theta}_n + t/\sqrt{n}) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right)}{\int_{\mathbb{R}} \pi(\hat{\theta}_n + u/\sqrt{n}) \exp\left(l_n(\hat{\theta}_n + u/\sqrt{n}) - l_n(\hat{\theta}_n)\right) du},\end{aligned}$$

for all $t \in \mathbb{R}$. We will complete the proof in the next lecture. □

22 Lecture 13/4

22.1 Proof of BvM (cont.)

Recall the statement of the Bernstein-von Mises (BvM) theorem: Let $Y_1, \ldots, Y_n \overset{iid}{\sim} p^\star$. Suppose Θ_n is a random variable that follows the posterior distribution given n observations and $\hat{\theta}_n$ is the MLE. Let π_n be the density of $T_n = \sqrt{n}(\Theta_n - \hat{\theta}_n)$. Then under regularity conditions (defined later),

$$\int_{\mathcal{H}} \left| \pi_n(t) - \frac{1}{\sqrt{(2\pi)^k |V^\star|}} \exp\left[-\frac{1}{2} t^\mathsf{T} V^{\star -1} t\right] \right| dt \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0, \tag{56}$$

where k is the dimension of the parameter space $\mathcal{H}$,

$$V^\star = \left[\mathbb{E}_\star \left(-\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta^\star}\right)\right]^{-1},$$

and $\theta^\star$ is the unique maximiser of $\theta \mapsto \mathbb{E}_\star\left[\log p_\theta(Y)\right]$.

(56) is the total variation distance between π_n and an MVN. So the BvM theorem is saying that the posterior is going to a Gaussian random variable in TV.

Proof cont. From last lecture,

$$\pi_n(t) = \frac{\pi(\hat{\theta}_n + t/\sqrt{n}) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right)}{\int_{\mathbb{R}} \pi(\hat{\theta}_n + u/\sqrt{n}) \exp\left(l_n(\hat{\theta}_n + u/\sqrt{n}) - l_n(\hat{\theta}_n)\right) du}. \tag{57}$$

Let C_n be the denominator.

To show (56), it suffices to show that

$$I = \int_{\mathbb{R}} \left| \pi\left(\hat{\theta}_n + t/\sqrt{n}\right) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right) - \pi(\theta^\star) \exp\left(-\frac{1}{2V^\star} t^2\right) \right| dt \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0. \tag{58}$$

Why? Assuming (58), we get

$$C_n \xrightarrow[n\to\infty]{p^\star\text{-prob.}} \pi(\theta^\star)\sqrt{2\pi V^\star},$$

since

$$\left| C_n - \pi(\theta^\star) \int_{\mathbb{R}} \exp\left(-\frac{1}{2V^\star} t^2\right) \right| \leq I.$$

Further,

$$\begin{aligned}
&\int_{\mathcal{R}} \left| \pi_n(t) - \frac{1}{\sqrt{(2\pi)^k |V^\star|}} \exp\left[-\frac{1}{2} t^\intercal V^{\star-1} t\right] \right| dt \\
&\quad \leq \frac{1}{C_n} \int_{\mathbb{R}} \left| \pi(\hat{\theta}_n + t/\sqrt{n}) \exp\left[l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right] - C_n \frac{1}{\sqrt{2\pi V^\star}} \exp\left(-\frac{1}{2V^\star} t^2\right) \right| dt \\
&\quad \leq \frac{1}{C_n} I + \frac{1}{C_n} \int_{\mathbb{R}} \left| \pi(\theta^\star) \exp\left(-\frac{1}{2V^\star} t^2\right) - C_n \frac{\exp\left(-\frac{1}{2V^\star} t^2\right)}{\sqrt{2\pi V^\star}} \right| dt,
\end{aligned}$$

where the second line follows by (57). Both of the terms on the RHS go to 0 in $p^\star$-probability if (58) holds.

We now shift our attention to proving (58). Assume that $\theta^\star$ is the unique maximiser of $l^\star$ and that the required conditions hold so that $\hat{\theta}_n \xrightarrow{p^\star} \theta^\star$, a.s. Define $h_n = -\frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 \log p_{\hat{\theta}_n}(Y_i)$ and assume

$$h_n \xrightarrow{p^\star\text{-a.s.}} V^{\star-1} 0.$$

First we will show that

$$\int_{\mathbb{R}} \left| \pi(\hat{\theta}_n) \exp\left(-\frac{h_n}{2}t^2\right) - \pi(\theta^\star) \exp\left(-\frac{1}{2V^\star}t^2\right) \right| dt \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0. \tag{59}$$

How will we do this? Assume π is continuous so that $\pi(\hat{\theta}_n) \to \pi(\theta^\star)$ a.s. Use the dominated convergence theorem with dominator

$$\exp\left(-\frac{h_n}{2}t^2\right) \leq \exp\left(-\frac{c}{2}t^2\right),$$

for some c and large enough n. (This holds since $V^{\star-1} > 0$ by assumption, so that h_n is positive eventually.)

Now the idea is that in I, replace $\pi(\theta^\star)\left(-\frac{1}{2V^\star}t^2\right)$ with $\pi(\hat{\theta}_n)\exp\left(-\frac{h_n}{2}t^2\right)$. (59) implies that (58) holds if

$$\int_{\mathbb{R}} \left| \pi\left(\hat{\theta}_n + t/\sqrt{n}\right) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n) \exp\left(-\frac{h_n}{2}t^2\right) \right| dt \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0 \tag{60}$$

Split the integral in (60) into two parts:

$$\begin{aligned} I_1 &= \int_{|t|>\delta\sqrt{n}} \left| \pi\left(\hat{\theta}_n + t/\sqrt{n}\right) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n) \exp\left(-\frac{h_n}{2}t^2\right) \right| dt, \\ I_2 &= \int_{|t|\leq\delta\sqrt{n}} \left| \pi\left(\hat{\theta}_n + t/\sqrt{n}\right) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n) \exp\left(-\frac{h_n}{2}t^2\right) \right| dt. \end{aligned}$$

By the triangle inequality,

$$\begin{aligned} I_1 &\leq \int_{|t|>\delta\sqrt{n}} \pi\left(\hat{\theta}_n + t/\sqrt{n}\right) \exp\left(l_n(\hat{\theta}_n + t/\sqrt{n}) - l_n(\hat{\theta}_n)\right) dt \\ &\quad + \int_{|t|>\delta\sqrt{n}} \pi(\hat{\theta}_n) \exp\left(-\frac{h_n}{2}t^2\right) dt \\ &= I_{11} + I_{12}. \end{aligned}$$

First we show that $I_{12} \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0$: For sufficiently large n (so that $h_n > 0$, $p^\star$-a.s.),

$$I_{12} = \pi(\hat{\theta}_n)\sqrt{\frac{2\pi}{h_n}} \int_{|t|>\delta\sqrt{n}} \sqrt{\frac{h_n}{2\pi}} \exp\left(-\frac{h_n}{2}t^2\right) dt$$

$$\leq \frac{\pi(\hat{\theta}_n)}{\delta\sqrt{n}h_n} \exp\left(-\frac{\delta^2 n h_n}{2}\right)$$
$$\xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0,$$

where the second line follows using the inequality $\mathbb{P}(Z > z) \leq \frac{\exp\left(-z^2/2\right)}{2\sqrt{2\pi}}$, for $Z \sim \mathcal{N}(0,1)$; and the third line follows since $h_n \to V^{\star-1} > 0$ and $\pi(\hat{\theta}_n) \xrightarrow{p^\star\text{-a.s.}} \pi(\theta^\star)$. Next we prove that $I_{11} \xrightarrow[n\to\infty]{p^\star\text{-prob.}} 0$: Assume that, with $p^\star$ probability going to 1,

$$\sup_{|u|>\delta} \frac{1}{n} l_n(\hat{\theta}_n + u) < \frac{1}{n} l_n(\hat{\theta}_n) - \eta,$$

for some $\eta = \eta(\delta) > 0$ which doesn't depend on n. (This looks a lot like the well-separate mode condition that we have seen earlier.) Then

$$I_{11} \leq \exp\left(\sup_{|u|>\delta}\left|l_n(\hat{\theta}_n + u) - l_n(\hat{\theta}_n)\right|\right) \int_{|t|>\delta\sqrt{n}} \pi(\hat{\theta}_n + t/\sqrt{n})dt$$
$$\leq \exp\left(-n\eta\right)\sqrt{n}\int_{|u|>\delta} \pi(\hat{\theta}_n + u)du$$
$$\to 0,$$

as $n \to \infty$ (surely), since the integral is bounded between zero and one, and $\eta > 0$.

Now we consider I_2: Here we can do a Taylor expansion. First do a change of variables:

$$I_2 = \sqrt{n}\int_{\{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta\}}\left|\pi(\theta)\exp\left(l_n(\theta) - l_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n)\exp\left(-\frac{h_n}{2}n(\theta-\hat{\theta}_n)^2\right)\right|d\theta.$$

Suppose that $\pi(\cdot)$ is continuous in a neighbourhood of $\theta^\star$ and that $\hat{\theta}_n \xrightarrow{p^\star\text{-a.s.}} \theta^\star$. Then for any $\epsilon > 0$, we can guarantee that

$$\pi(\theta) \in \left[\pi(\hat{\theta}_n)(1-\epsilon), \pi(\hat{\theta}_n)(1+\epsilon)\right],$$

for all θ satisfying $\left|\theta - \hat{\theta}_n\right| \leq \delta$, by picking δ small enough and n large enough.

Then

$$I_2 \leq \sqrt{n}\left(1+\epsilon\right)\pi(\hat{\theta}_n)\int_{\{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta\}}\left|\exp\left[l_n(\theta) - l_n(\hat{\theta}_n)\right] - \exp\left[-\frac{h_n}{2}n\left(\theta-\hat{\theta}_n\right)^2\right]\right|d\theta.$$

Now Taylor expand:

$$l_n(\theta) = l_n(\hat{\theta}_n) + \frac{\nabla^2_\theta l_n(\hat{\theta}_n)}{2}\left(\theta - \hat{\theta}_n\right)^2 + \frac{\nabla^2_\theta l_n(\tilde{\theta}_n) - \nabla^2_\theta l_n(\hat{\theta}_n)}{2}\left(\theta - \hat{\theta}_n\right)^2,$$

for all θ satisfying $\left|\theta - \hat{\theta}_n\right| \leq \delta$. Denote the third term on the RHS by $R_n(\theta)$. Then

$$l_n(\theta) - l_n(\hat{\theta}_n) = -\frac{nh_n}{2}\left(\theta - \hat{\theta}_n\right)^2 + R_n(\theta).$$

We will bound the remainder term $R_n(\theta)$. We need to assume that for all $\epsilon > 0$, we can find δ small enough such that

$$\sup_{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta}\left|\frac{\nabla^2_\theta l_n(\theta) - \nabla^2_\theta l_n(\hat{\theta}_n)}{n}\right| \leq \epsilon,$$

with $p^\star$-probability going to 1. Hence

$$\sup_{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta}|R_n(\theta)| \leq \frac{\epsilon n\left(\theta - \hat{\theta}_n\right)^2}{2}, \tag{61}$$

with $p^\star$-probability going to 1. With some algebra,

$$\begin{aligned}
I_2 &\leq \sqrt{n}(1+\epsilon)\pi(\hat{\theta}_n)\int_{\{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta\}} \exp\left[-\frac{nh_n}{2}(\theta-\hat{\theta}_n)^2\right]|\exp(R_n(\theta)) - 1|d\theta \\
&\leq \sqrt{n}(1+\epsilon)\pi(\hat{\theta}_n)\int_{\{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta\}} \exp\left[-\frac{nh_n}{2}(\theta-\hat{\theta}_n)^2\right]|R_n(\theta)|\exp|R_n(\theta)|d\theta \\
&\leq \frac{\epsilon n^{3/2}}{2}(1+\epsilon)\pi(\hat{\theta}_n)\int_{\{\theta:\left|\theta-\hat{\theta}_n\right|\leq\delta\}} \exp\left[-\frac{nc}{2}(\theta-\hat{\theta}_n)^2\right]\left(\theta - \hat{\theta}_n\right)^2 d\theta \\
&= \frac{\epsilon n^{3/2}}{2}(1+\epsilon)\pi(\hat{\theta}_n)\left[\frac{1}{nc}\sqrt{\frac{2\pi}{nc}}\right] \\
&= \frac{\epsilon(1+\epsilon)\pi(\hat{\theta}_n)\sqrt{2\pi}}{2c^{3/2}},
\end{aligned}$$

where the second line uses the inequality $|exp(x) - 1| \leq |x|\exp|x|$; the third uses the upper bound (61) on $R_n(\theta)$ and a lower bound $c > 0$ of h_n for large enough n; and the fourth follows since the integral on the third line is the variance of a Gaussian (up to the normalising constant).

Now, ϵ can be made arbitrarily small as $n \to \infty$. Hence we must have $I_2 \xrightarrow{p^\star\text{-prob.}} 0$. □

22.2 Conditions for the BvM theorem

We collect the assumptions used in the proof of the BvM theorem:

1. $\theta \mapsto l^\star(\theta) = \int_{\mathcal{Y}} \log p_\theta(y) p^\star(dy)$ is uniquely maximised at $\theta^\star$.
2. The MLE $\hat{\theta}_n$ converges $p^\star$-a.s. to $\theta^\star$. (This can be weakened to $p^\star$-probability convergence without changing the proof.)
3. There exists a neighbourhood U of $\theta^\star$ such that $p_\theta(y) > 0$, $y \mapsto p_\theta(y)$ is measurable and $\theta \mapsto p_\theta(y)$ is twice continuously differentiable, for all $y \in \mathcal{Y}$ and $\theta \in U$.
4. $h_n = -\frac{1}{n}\nabla_\theta^2 \log p_{\hat{\theta}_n}(Y_i) \xrightarrow{p^\star\text{-prob.}} V^{\star-1} > 0$.
5. The prior π is continuous and positive in a neighbourhood of $\theta^\star$.
6. For any $\delta > 0, \exists \eta > 0$ such that
$$\mathbb{P}_\star \left[\sup_{|u|>\delta} \frac{1}{n} l_n(\hat{\theta}_n + u) < \frac{1}{n} l_n(\hat{\theta}_n) - \eta \right] \xrightarrow{n\to\infty} 1.$$
7. Finally, for any $\epsilon > 0 \exists \delta > 0$, such that
$$\sup_{\theta: \left|\theta - \hat{\theta}_n\right| \leq \delta} \left| \frac{\nabla_\theta^2 l_n(\theta) - \nabla_\theta^2 l_n(\hat{\theta}_n)}{n} \right| \leq \epsilon,$$
with $p^\star$-probability going to 1.

22.3 Bayesian point estimation

Theorem 22.1 (Theorem 4.4 of [GR]). *Under the BvM conditions above, along with the assumption $\int |\theta|\pi(\theta)d\theta < \infty$, the posterior mean $\mathbb{E}\left[\Theta|Y_1,\ldots,Y_n\right]$ satisfies*

$$\sqrt{n}\left(\hat{\theta}_n - \mathbb{E}\left[\Theta|Y_1,\ldots,Y_n\right]\right) \xrightarrow{p^\star\text{-}prob.} 0.$$

Why is this interesting? The theorem implies that the distribution of the posterior mean is approximately equal to the distribution of the MLE. This means that the asymptotic variance of the posterior mean is given by the sandwich formula for the MLE in misspecified models. This is in spite of the fact that $V^\star$ in BvM – which is kind of like the asymptotic variance of the posterior – isn't given by the sandwich formula (as we showed earlier).

References

[AKJ20] A. E. Alaoui, F. Krzakala, and M. Jordan. Fundamental limits of detection in the spiked Wigner model. *The Annals of Statistics*, 48(2):863–885, Apr. 2020. doi:10.1214/19-AOS1826.

[Bil12] P. Billingsley. *Probability and Measure.* Wiley Series in Probability and Statistics. Wiley, Hoboken, N.J, anniversary ed edition, 2012.

[BM18] D. Banerjee and Z. Ma. Asymptotic normality and analysis of variance of log-likelihood ratios in spiked random matrix models. *arXiv:1804.00567 [cs, math, stat]*, Apr. 2018, 1804.00567.

[Cat11] A. Caticha. Entropic Inference. *arXiv:1011.0723 [cond-mat, physics:physics, stat]*, pages 20–29, 2011, 1011.0723. doi:10.1063/1.3573619.

[CCD$^+$18] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, Feb. 2018. doi:10.1111/ectj.12097.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, Oct. 2019, 1809.09573. doi:10.1109/TSP.2019.2937282.

[EKBB$^+$13] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, Sept. 2013. doi:10.1073/pnas.1307842110.

[FWCT] C. Feng, H. Wang, T. Chen, and X. M. Tu. On exact forms of Taylor's theorem for vector-valued functions. *Biometrika*, 101(4):1003–1003, 2014.

[GR] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer, New York, 2003.

[HL20] H. Hu and Y. M. Lu. Universality Laws for High-Dimensional Learning with Random Features. *arXiv:2009.07669 [cs, math]*, Sept. 2020, 2009.07669.

[JO20] I. M. Johnstone and A. Onatski. Testing in high-dimensional spiked models. *The Annals of Statistics*, 48(3):1231–1254, June 2020. doi:10.1214/18-AOS1697.

[LC06] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, May 2006.

[LS20] C. Lu and S. Sen. Contextual Stochastic Block Model: Sharp Thresholds and Contiguity. *arXiv:2011.09841 [cs, math, stat]*, Nov. 2020, 2011.09841.

[Rou16] J. Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. *Annual Review of Statistics and Its Application*, 3(1):211–231, 2016. doi:10.1146/annurev-statistics-041715-033523.

[SC18] P. Sur and E. J. Candes. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv:1803.06964 [math, stat]*, June 2018, 1803.06964.

[TSH] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

[vdV] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi:10.1017/CBO9780511802256.