1 Going beyond ML with ML type methods

1.1 When ML fails

Example. Is LSAT correlated with GPA? Suppose we want to answer this question based on data \((X_i, Y_i), i = 1, \ldots, n\), on LSAT and GPA scores of \(n\) students. As in Hw 1, we could model \((X_i, Y_i) \overset{iid}{\sim} \text{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)\) and formalize the test on the correlation coefficient parameter \(\rho\). However, there is no compelling reason to believe that the two scores are jointly normally distributed. A more general model and test formalization, with almost no assumptions on the shape of the bivariate pdf is given by

\[
(X_i, Y_i) \overset{\text{iid}}{\sim} g, g \in G_2 = \{\text{all densities on } \mathbb{R}^2 \text{ with finite second moments}\}
\]

test \(H_0 : \rho(g) = 0\) vs. \(H_1 : \rho(g) \neq 0\)

where\(^1\)

\[
\rho(g) = \text{Corr}(X, Y | g)
\]

Unfortunately, the ML approach does not work here because \(\hat{g}_{\text{MLE}}\) does not exist! Indeed, for any observed data \((x_i, y_i), i = 1, \ldots, n\) we have

\[
\sup_{g \in G_2} \ell_{x,y}(g) = \infty. \tag{1}
\]

As a quick check of (1), define for every \(k \geq 1\),

\[
g_k := \frac{1}{n} \sum_{i=1}^{n} N_2(z_i, I_2) \in G_2, \text{ where } z_i = (x_i, y_i)^T
\]

and notice that for \(k\) large, \(\ell_{x,y}(g_k) \approx n \log k \uparrow \infty\) as \(k \to \infty\).

1.2 Methods based on M-estimators

Consider the situation where data \(X_1, \ldots, X_n\) are modeled as \(X_i \overset{\text{iid}}{\sim} g, g \in G\) and we are interested in the parameter \(\theta = \theta(g) \in \Theta \subset \mathbb{R}^d\). Here \(\theta\) provides only a partial indexing of the model space \(G\). We could write \(G = \cup_{a \in \Theta} \{g \in G : \theta(g) = a\}\).

An ML type approach for inference on \(\theta\) could be developed by starting with some “criterion” function \(m(X_i, \theta)\) with the property that large values of \(m(X_i, \theta)\) indicate a closer agreement\(^2\) between \(\theta\) and \(X_i\). This could be made more precise in the following way: if the

\[1\]More precisely, \(\rho(g) = \frac{\sigma_{12}(g)}{\sigma_1(g)\sigma_2(g)}\) with \(\sigma_{12}(g) = \int \int (x - \mu_1(g))(y - \mu_2(g))g(x, y)dx\,dy, \sigma_1^2(g) = \int \int (x - \mu_1(g))^2g(x, y)dx\,dy, \sigma_2^2(g) = \int \int (y - \mu_2(g))^2g(x, y)dy\,dx, \mu_1(g) = \int xg(x, y)dx\,dy, \mu_2(g) = \int yg(x, y)dx\,dy\)  

\[2\]The text uses “discrepancy measure” to mean the negative of a criterion function.
true pdf is $g = g_0 \in \mathcal{G}$ and so the true $\theta = \theta_0 := \theta(g_0)$, then the map $\theta \mapsto \text{E}\{m(X_i, \theta)|g_0\}$ is maximized at $\theta_0$ and is potentially a decreasing function $\|\theta - \theta_0\|$ at least in a small neighborhood of $\theta_0$.

Testing of $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \not\in \Theta_0$ could be carried out by the test that rejects $H_0$ for large values of the statistic

$$M_n = \max_{\theta \in \Theta} m_n(\theta) - \max_{\theta \in \Theta_0} m_n(\theta)$$

where

$$m_n(\theta) = \sum_{i=1}^{n} m(X_i, \theta).$$

Under reasonable regularity conditions we might expect $m_n(\theta)$ to be approximately quadratic around a neighborhood of its maxima $\hat{\theta}_n = \arg\max_{\theta \in \Theta} m_n(\theta)$ with curvature $\{-\bar{m}_n(\hat{\theta}_n)\}$. Hence the above test will be approximately equivalent to reject $H_0$ whenever

$$\{\theta : (\theta - \hat{\theta}_n)^T \{-\bar{m}_n(\hat{\theta}_n)\}(\theta - \hat{\theta}_n) \leq c^2 \} \cap \Theta_0 = \emptyset$$

for some $c$. The maxima $\hat{\theta}_n$, which can be taken as an “estimator” of $\theta$, is commonly referred to as an M-estimator (M for maximization).

In case of a parametric model where $\theta$ provides a complete indexing, i.e., we could rewrite the mode as $X_i \overset{\text{iid}}{\sim} g(x_i|\theta)$, an attractive criterion function is $m(X_i, \theta) = \log g(x_i|\theta)$, with $\text{E}\{m(X_i, \theta)|\theta_0\} = \text{const} - K(g(\cdot|\theta_0), g(\cdot|\theta))$, taking us back to the ML approach. However, other criterion functions could be used in this situation and will have to be used in other situations where $\theta$ provides only a partial indexing. Here are some examples.

**Example** (Least squares, linear and non-linear). Consider data $X_i = (Z_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \cdots, n$ modeled as

$$Y_i = h(Z_i, \beta) + \sigma \epsilon_i, \quad (Z_i, \epsilon_i) \overset{\text{iid}}{\sim} g(z, \epsilon)$$

with model parameters $(\beta, \sigma, g) \in \mathbb{R}^p \times \mathbb{R}_+ \times \mathcal{G}$, where $\mathcal{G}$ is the space of all pdfs $g(z, \epsilon)$ on $\mathbb{R}^p \times \mathbb{R}$ such that $\text{E}[\epsilon|Z, g] = 0$, $\text{Var}[\epsilon|Z, g] = 1$. Here $h$ is a known function, potentially non-linear and satisfies the identifiability condition:

$$h(z, \beta_1) = h(z, \beta_2) \quad \forall z \iff \beta_1 = \beta_2.$$ 

The least squares criterion function is: $m(X_1, \beta) = -(Y_1 - h(Z_1, \beta))^2$. If $(\beta_0, \sigma_0, g_0)$ characterizes the true distribution of the data, and $E_0$ denotes expectation under this distribution, then

$$\text{E}_0\{m(X_1, \beta)\} = -\sigma_0^2 - \text{E}_0\{h(Z_1, \beta_0) - h(Z_1, \beta)\}^2$$

which is uniquely maximized at $\beta = \beta_0$. The M-estimator $\hat{\beta}_n$ is often computed by using quasi-Newton optimization algorithms.

**Example** (Quantile regression). Consider data $X_i = (Z_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \cdots, n$ and suppose we want to model the $\tau$-th conditional quantile of $Y_i$ given $Z_i$ by $Z_i^T \beta$. More formally,

$$Y_i = Z_i^T \beta + \epsilon_i, \quad Z_i \overset{\text{iid}}{\sim} g_z(z_i), \quad \epsilon_i|Z_i \overset{\text{iid}}{\sim} g_\epsilon(\epsilon_i|Z_i)$$
where \((\beta, g_z, g_t)\) are unknown but \(g_t\) is constrained to satisfy: \(\int_{-\infty}^{0} g_t(\varepsilon | z_i) d\varepsilon = \tau\), so that under the model \(P(Y_i \leq z_i^T \beta | Z_i = z_i) = \tau\) for all \(z_i\). Koenker and Bassett introduced the criterion function

\[
m(X_i, \beta) = -(Y_i - Z_i^T \beta) \{\tau - I(Y_i - Z_i^T \beta \leq 0)\}
\]

and it can be proved that \(E_0\{m(X_i, \beta)\}\) is uniquely maximized at \(\beta = \beta_0\). The M-estimator \(^\beta_n\) can be computed efficiently by linear programming.

### 1.3 \textbf{Z-estimators}

If the criterion function is differentiable in \(\theta\), then the M-estimator can potentially be computed by solving the equation:

\[
\psi_n(\theta) := \sum_{i=1}^{n} \psi(X_i, \theta) = 0
\]

where \(\psi(X_i, \theta) = \frac{\partial}{\partial \theta} m(X_i, \theta)\). This suggests another way to get to a good estimator of \(\theta\), which can then be used to construct test procedures for hypotheses about \(\theta\). Let \(\dim(\theta) = d\). We will call a \(d\)-dimensional map \(\theta \mapsto \psi(X_i, \theta)\) a score function if \(E_0\psi(X_i, \theta)\) has a unique zero at \(\theta = \theta_0\). A root (or zero) \(^\theta_n\) of \(\psi_n(\theta)\) will be called a Z-estimator. Bear in mind that an M-estimator need not be a Z-estimator (MLE for the model \(X_i \sim \text{Uniform}(0, \theta)\)) and a Z-estimator need not be an M-estimator (moment estimators, next).

**Example** (Method of Moments). Consider the model \(X_1, \cdots, X_n \sim g(x_i | \theta), \theta \in \Theta \subset \mathbb{R}^d\). Suppose \(X_j\)'s are univariate. Let

\[
\mu_j(\theta) = E\{X_i^j| \theta\}, j = 1, \cdots, d.
\]

Suppose \(\theta \mapsto (\mu_1(\theta), \mu_2(\theta), \cdots, \mu_d(\theta))\) is one-to-one. Then one could estimate \(\theta\) by solving the equations:

\[
\mu_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i^j, \quad j = 1, 2, \cdots, d
\]

The solution \(^\theta\) is called the method of moments (MoM) estimator. Similar methods could be devised for multivariate \(X_j\)'s.

### 1.4 \textbf{Asymptotic Normality}

Under regularity conditions, asymptotic normality properties could be established for both M- and Z-estimators, leading to asymptotic size and p-value calculations for tests based on these estimators. The classical conditions are precisely the Cramér conditions we saw for asymptotic normality of the MLE. In fact we can reproduce Theorem 7 from Handout 3 almost verbatim. Here is a precise statement for the Z-estimator (and so also applies to M-estimators that can be seen as Z-estimators).
Theorem 1. Suppose $X_1, X_2, \ldots$ are modeled as $X_i \sim g$, $g \in G$ and we are interested in $\theta = \theta(g) \in \Theta \subset \mathbb{R}^d$. Assume $g_0 \in G$ is the true distribution and $\theta_0 = \theta(g_0)$ is the true value of $\theta$. Let $E_0$, $\text{Var}_0$ denote expectation and variance under $g_0$. Suppose $\theta \mapsto \psi(x, \theta)$ is twice continuously differentiable in a neighborhood $U_0$ of $\theta_0$ for every $x$. Suppose $E_0\psi(x, \theta_0) = 0$, $\text{Var}_0\psi(X_i, \theta_0)$ is finite and the matrix $E_0\psi(X_i, \theta_0)$ exists and is nonsingular. Assume that the second order partial derivatives are dominated within $U_0$ by a fixed function $\psi(x)$ with $E_0\psi(X_1) < \infty$. Then,

1. The probability that the equation $\psi_n(\theta) = 0$ has at least one root tends to 1 as $n \to \infty$ and there exists a sequence of roots $\hat{\theta}_n$ such that $\hat{\theta}_n \overset{p}{\to} \theta_0$.

2. For any consistent estimator sequence $\hat{\theta}_n$ satisfying $\psi_n(\hat{\theta}_n) = 0$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left\{E_0\psi(X_1, \theta_0)^T\right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i; \theta_0) + R_n$$

with $R_n \overset{p}{\to} 0$, and hence,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N_d(0, \Sigma(g_0))$$

where $\Sigma(g_0) = \left\{E_0\psi(X_1, \theta_0)^T\right\}^{-1} \left\{\text{Var}_0\psi(X_1, \theta_0)\right\} \left\{E_0\psi(X_1, \theta_0)\right\}^{-1}$.

3. With probability tending to 1, $\frac{1}{n}\psi_n(\hat{\theta}_n)$ is nonsingular and

$$\tilde{\Sigma}_n := \left\{\frac{1}{n}\psi_n(\hat{\theta}_n)^T\right\}^{-1} \left\{\frac{1}{n} \sum_{i=1}^n \psi(X_i; \hat{\theta}_n)\psi(X_i; \hat{\theta}_n)^T\right\} \left\{\frac{1}{n}\psi_n(\hat{\theta}_n)\right\}^{-1} \overset{p}{\to} \Sigma(g_0)$$

and hence, \(\hat{\theta}_n \sim AN_d(\theta_0, \frac{1}{n}\tilde{\Sigma}_n)\).

Proof. Repeat the proof of Theorem 7 of Handout 3 by replacing $\hat{\ell}_n(\theta)$ with $\psi_n(\theta)$. 

Example (Error-in-variable regression). Let observations $X_i = (Z_i, Y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, 2, \ldots$ denote paired univariate measurements modeled as

$$Z_i = U_i + \phi_i, \quad Y_i = \alpha + \beta U_i + \epsilon_i, \quad (U_i, \phi_i, \epsilon_i) \overset{\text{iid}}{\sim} g_U \times N(0, \sigma^2) \times N(0, \sigma^2).$$

This model is referred to as error-in-variable regression or measurement-error regression, since the actual predictor $U_i$ is unobserved, but we get to record a contaminated version $Z_i = U_i + \phi_i$. The distribution $g_U$ of the latent predictor is unknown, and is of no interest. Primary interest concerns regression coefficients $\theta = (\alpha, \beta, \sigma^2)$. The simple least squares estimate of $\theta$ provides poor inference, the estimator is not even consistent. But a consistent, asymptotically normal Z-estimator is obtained by using (HW 2):

$$\psi(X_i, \theta) = \begin{pmatrix} Y_i - \alpha - \beta Z_i \\ Z_i(Y_i - \alpha - \beta Z_i) + \beta \sigma^2 \\ (Y_i - \alpha - \beta Z_i)^2 - \sigma^2(1 + \beta^2) \end{pmatrix}.$$ 

\(^3\text{Here we have assumed } \text{Var}(\phi_i) = \text{Var}(\epsilon_i) = \sigma^2. \text{ In a more general setting we could let } \text{Var}(\phi_i) = \rho \sigma^2. \text{ When } \rho \text{ is known, we could get back to the equal variance setting by rescaling } U_i \text{ and } Z_i. \text{ Unknown } \rho \text{ is difficult to deal with, and requires prior knowledge or additional variables.} \)
2 Other approaches

The M- and Z-estimator approaches provide a broad platform to construct statistical procedures with sound frequentist properties. They cover a large number of widely used statistical methods. However, there are an even larger number of procedures that do not stem from a particular platform of methods generation, but are constructed more or less based on heuristics that apply to a particular context. A testing procedure built upon heuristics, turns into a rigorous testing procedure the moment we are able to calculate its size and establish good power on the alternatives. These calculations can be challenging, but asymptotic theory and/or simulation techniques help out. Here are two specific examples.

2.1 Rank based test for location shift

Consider two samples of observations \( W_1, \ldots, W_m \) and \( V_1, \ldots, V_k \) modeled as 

\[
W_i \sim g(w_i - \theta), \quad V_j \sim g(v_j), \quad W_i's, V_j's \text{ independent}
\]

where \( \theta \in \mathbb{R} \) and \( g \in \{ \text{all densities on } \mathbb{R} \} \) are unknown. We are interested in testing \( H_0 : \theta = 0 \), vs. \( H_1 : \theta \neq 0 \). A popular test statistic is the Mann-Whitney statistic\(^4\):

\[
U = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} I(V_j \leq W_i)
\]

Clearly, \( \mu(\theta) := \mathbb{E}(U|\theta) = P(V_1 \leq W_1|\theta) \), which equals 1/2 if \( \theta = 0 \). If \( \theta > 0 \), then \( U \) is expected to be larger (because then \( W_i \)'s get shifted to right by \( \theta \) and should be relatively larger than \( V_j \)'s), and similarly, for \( \theta < 0 \), \( U \) is expected to be smaller. Hence a test statistic against \( H_0 \) is given by \( |U - 1/2| \).

It can be shown (from U-statistic theory, beyond the scope of this course) that with \( n = m + k \) and \( G(w) \) denoting the CDF of \( g(w) \),

\[
\sqrt{n}(U - \mu(\theta)) = \frac{\sqrt{n}}{m} \sum_{i=1}^{m} [G(W_i) - \mathbb{E}\{G(W_i)|\theta]\} - \frac{\sqrt{n}}{k} \sum_{j=1}^{k} [G(V_j - \theta) - \mathbb{E}\{G(V_j - \theta)|\theta]\} + R_n(\theta)
\]

with \( R_n(\theta) \xrightarrow{P} 0 \) under \( \theta \) as \( n \to \infty \). If \( m/n \to \lambda \in (0, 1) \) as \( n \to \infty \), then we can conclude from the above (by standard CLT) that

\[
\sqrt{n}(U - \mu(\theta)) \xrightarrow{D} N_1(0, \sigma^2(\theta))
\]

where

\[
\sigma^2(\theta) = \frac{1}{\lambda} \text{Var}\{G(W_1)|\theta}\} + \frac{1}{1-\lambda} \text{Var}\{G(V_1 - \theta)|\theta}\}.
\]

\(^4\)A related statistic is the Wilcoxon rank sum statistic: \( R = \sum_{i=1}^{m} \text{rank}(W_i) = mkU + m(m + 1)/2 \) where \( \text{rank}(W_i) \) gives the rank of \( W_i \) in the pooled sample \( \{W_1, \ldots, W_m, V_1, \ldots, V_k\} \) arranged from smallest to largest.
These calculations simplify considerably when $\theta = 0$, which is what we need to calculate the size of the test based on $U$. Note that when $\theta = 0$, $W_i, V_j \overset{\text{iid}}{\sim} g$ and hence $G(W_i)$ and $G(V_j)$ are all iid Uniform$(0, 1)$ variables with variance $1/12$. Hence,

$$\sigma^2(0) = \frac{1}{12\lambda(1 - \lambda)}.$$

A simple consistent estimator of $\sigma^2(0)$ is $n^2/\{12mk\}$ and hence an approximately size-$\alpha$ Mann-Whitney test for $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$ is given by “reject $H_0$ if $\sqrt{12mk|U - 1/2|/n} \geq \Phi^{-1}(1 - \alpha/2)$”.

Of course, if $g$ was known to be a Gaussian, then the ML test for the above hypotheses would be given by the standard t-test (with equal variance assumed). We will later see that even in this case, a size-$\alpha$ Mann-Whitney test has power comparable to a size-$\alpha$ t-test. Much more interestingly, if $g$ is non-Gaussian and possesses tails that are thicker than those of a Gaussian pdf, then the Mann-Whitney test could be much more powerful than the t-test!

2.2 Testing complete spatial randomness (CSR)

Point pattern data are routinely studied in ecology and related fields. Let $X_1, \ldots, X_n$ denote observed spatial locations of $n$ events (such as occurrences of some shrub species) in $S \subset \mathbb{R}^2$. Can we detect whether the point pattern of these locations is aggregate (i.e., points are clustered together), or regular (i.e., grid-like and the points repel each other) or completely spatially random?

A heuristic approach to this inference problem goes like this. The “$K$-statistic” at distance $r$ is defined as

$$\hat{K}(r) = \frac{A}{n^2} \sum_{i \neq j} w_{ij} I(\|X_i - X_j\| \leq r)$$

where $A = \text{area}(S)$, $w_{ij} = 1/\text{area}(S \cap B(X_i, r))$ with $B(c, r) := \{x \in \mathbb{R}^2 : \|x - c\| \leq r\}$ denoting the ball/disc in $\mathbb{R}^2$ with center $c$ and radius $r$. The hypothesis of CSR implies $K(r) \approx \pi r^2$. So we will take our test statistic as $\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} - r$

The exact distribution of $\hat{L}(r)$ is difficult to assess (depends on the shape of $S$ in an intractable way). But it is easy to simulate CSR patterns of $n$ points in $S$, and we could use the simulations to create Monte Carlo approximation to the distribution of $\hat{L}(r)$. If $L_{\alpha/2, r}$ and $L_{1-\alpha/2, r}$ denote the (MC approximate) $\alpha/2$ and $(1 - \alpha/2)$-th quantiles of $L(r)$ under CSR, then we could reject “$H_0 : \text{CSR at range } r$” if $\hat{L}(r) \notin [L_{\alpha/2, r}, L_{1-\alpha/2, r}]$. See http://onlinelibrary.wiley.com/doi/10.1111/j.0906-7590.2008.05374.x/abstract for more details and http://www.springerlink.com/content/j353r2728n1301v5/ for an application to shrub patterns.

3 Relative Efficiency of Tests

Now that we have seen multiple ways of constructing test procedures, it is natural to ask if we could rank them based on their relative performances. While such comparisons are hard to make for a fixed sample size, asymptotic calculations often help.
To facilitate our discussion, let’s get a clear formalization of the asymptotic context. We will consider an infinite sequence of statistical models \( X^{(n)} \sim p_{n,\theta}, \theta \in \Theta \) with a common parameter. For IID data \( X_1, X_2, \ldots \), modeled as \( X_i \overset{iid}{\sim} g(x_i|\theta), \theta \in \Theta \), such a sequence of models is given by

\[
X^{(n)} = (X_1, \ldots, X_n), \quad p_{n,\theta}(x^{(n)}) = g(x_1|\theta) \times \cdots \times g(x_n|\theta).
\] (4)

For testing \( H_0 : \theta \in \Theta_0 \) vs. \( H_1 : \theta \in \Theta_0^c \), we will consider sequences of test procedures given by test statistics and critical values, \( \{(T_n, c_n) : n = 1, 2, \ldots \} \) with \( T^{(n)} \) based on data \( X^{(n)} \), typically given by the same test generation scheme applied to all \( n \), such as ML, or a Z-estimation associated with a certain score function \( \psi \), etc.

Such a test sequence will be called asymptotically size \( \alpha \) if,

\[
\lim_{n \to \infty} \sup_{\theta \in \Theta_0} P(T_n \geq c_n|\theta) = \alpha.
\]

Now consider two sequences of test procedures sequences \( (T_{n,1}, c_{n,1})_n, (T_{n,2}, c_{n,2})_n \), \( n = 1, 2, \ldots \) and denote the corresponding power function sequences as:

\[
\pi_{n,1}(\theta) = P(T_{n,1} \geq c_{n,1}|\theta), \quad \pi_{n,2}(\theta) = P(T_{n,2} \geq c_{n,2}|\theta).
\]

If both test sequences are asymptotically size \( \alpha \), then we would declare the first sequence asymptotically better than the second if

\[
\lim_{n \to \infty} \pi_{n,1}(\theta) \geq \lim_{n \to \infty} \pi_{n,2}(\theta) \quad \text{for every } \theta \in \Theta_0^c.
\]

However, such comparisons are mostly useless because almost for all reasonable size \( \alpha \) test sequences, the limiting power equals 1 at any \( \theta \in \Theta_0^c \). For example, for the sequence of models based on \( X_i \overset{iid}{\sim} N(\mu, 1) \), an asymptotically size 5% test for \( H_0 : \mu \leq 0 \) vs \( H_1 : \mu \geq 0 \) is given by “reject \( H_0 \) if \( \sqrt{n}X^{(n)} \geq 1.64 \).” At any \( \mu > 0 \), \( \pi_n(\mu) = 1 - \Phi(1.64 - \sqrt{n}\mu) \to 1 \) as \( n \to \infty \).

### 3.1 The Pitman Relative Efficiency

The situation can be salvaged by comparing \( \lim_{n \to \infty} \pi_n(\theta_n) \) over a sequence of increasingly difficult alternatives \( \theta_n \to \partial \Theta_0 \) [the boundary of \( \Theta_0 \)]. To make headway along this line, we make a number of simplifying assumptions. First, we assume \( \Theta_0 = \{\theta_0\} \), a point null with \( \theta_0 \). We also focus only on test statistic sequences \( (T_n : n = 1, 2, \ldots) \) that are “regular” at \( \theta_0 \), i.e., there exist functions \( \mu(\theta) \) and \( \sigma(\theta) \) such that for any \( h \),

\[
\frac{\sqrt{n}(T_n - \mu(\theta_0 + \frac{h}{\sqrt{n}}))}{\sigma(\theta_0 + \frac{h}{\sqrt{n}})} \overset{d}{\to} N(0, 1) \quad \text{under } p_{n,\theta_0 + \frac{h}{\sqrt{n}}}.
\] (5)

The above condition is to be understood as follows: if \( Z_n \) denotes the normalized statistic on the left of (5) then \( \lim_{n \to \infty} P(Z_n \in A|\theta_0 + \frac{h}{\sqrt{n}}) = P(Z \in A) = P(Z \in A) \) for every \( A \subset \mathbb{R} \) where \( Z \sim N(0, 1) \). Bear in mind that regularity is a strong condition and need not hold even if
we could show $\sqrt{n}(T_n - \mu(\theta))/\sigma(\theta) \overset{\mathcal{D}}{\rightarrow} N(0, 1)$ at every $\theta \in \Theta$. We will later see tools to establish regularity.

As a sequence of increasingly difficult alternatives, consider $\theta_n = \theta_0 + h/\sqrt{n}$, $n \geq 1$, for any fixed $h \neq 0$, i.e., we consider the limiting power along a sequence of alternatives collapsing onto $\theta_0$ at $\sqrt{n}$ rate along a ray $h$. If $\{(T_n, c_n), n \geq 1\}$ is asymptotically size $\alpha$ with $T_n$ regular, then we can write $c_n = \mu(\theta_0) + \sigma(\theta_0)(z_\alpha + r_n)/\sqrt{n}$ where $z_\alpha = \Phi^{-1}(1 - \alpha)$ and $r_n \to 0$.

$$\lim_{n \to \infty} \pi_n(\theta_n) = \lim_{n \to \infty} P \left( \frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \geq \frac{\sqrt{n} \{\mu(\theta_0) + \sigma(\theta_0)(z_\alpha + r_n)/\sqrt{n} - \mu(\theta_n)\}}{\sigma(\theta_n)} \right).$$

Notice that,

$$\lim_{n \to \infty} \frac{\sqrt{n} \{\mu(\theta_0) + \sigma(\theta_0)(z_\alpha + r_n)/\sqrt{n} - \mu(\theta_n)\}}{\sigma(\theta_n)} = z_\alpha - \frac{h^T \hat{\mu}(\theta_0)}{\sigma(\theta_0)}$$

and hence by Slutsky's theorem,

$$\lim_{n \to \infty} \pi_n(\theta) = 1 - \Phi \left( z_\alpha - \frac{h^T \hat{\mu}(\theta_0)}{\sigma(\theta_0)} \right).$$

The quantity $\hat{\mu}(\theta_0)/\sigma(\theta_0)$ is known as the slope of the test statistic sequence $T_n$ at $\theta_0$. Clearly we should prefer test statistics with higher slopes!

The above calculations may seem too dependent on the arbitrary choice of the collapsing alternatives. Pitman established that these calculations are indeed robust. To formalize this, I state below the definition of "the Pitman relative efficiency" and a theorem that connects this concept with the ratio of slopes of two test sequences.

**Definition 1.** Let $\{T_{n,1} : n \geq 1\}$ and $\{T_{n,2} : n \geq 1\}$ be two test statistics sequences. For any given $\alpha, \gamma \in (0, 1)$, let $n_1(\alpha, \gamma, \theta)$ denote the minimum $n$ needed so that a level $\alpha$ test based on $T_{n,1}$ has power at least $\gamma$ at $\theta$. Let $n_2(\alpha, \gamma, \theta)$ denote the same for the other sequence. The Pitman relative efficiency of the first sequence against the second, at level $\alpha$, power $\gamma$ is defined as:

$$\lim_{\theta \to \theta_0} \frac{n_2(\alpha, \gamma, \theta)}{n_1(\alpha, \gamma, \theta)},$$

provided the limit exists.

In principle, the limiting sample size ratio may depend on $\alpha, \gamma$ and the particular collapsing sequence of alternatives $\theta_\nu$. But often it does not, especially if $\theta_\nu = \theta_0 + \nu h$ where $h$ is a fixed unit vector and $\nu \downarrow 0$, i.e., the collapsing alternatives $\theta_\nu$ approach $\theta_0$ along a ray and from only one direction. Below is a precise statement and proof for the case of $\dim(\theta) = 1$ with $\Theta = [0, \infty)$, $\theta_0 = 0$.

**Theorem 2.** Consider statistical models $\{p_{n,\theta} : \theta \geq 0\}$ such that $\|p_{n,\theta} - p_{n,0}\| \to 0$ as $\theta \to 0$ for every fixed $n$. Let $T_{n,1}$ and $T_{n,2}$ be two sequences of test statistics that satisfy the regularity condition (5) with $\hat{\mu}_i(0) > 0$ and $\sigma_i(0) > 0$, $i = 1, 2$. Then the Pitman relative efficiency of the first sequence to the second equals

$$\left( \frac{\hat{\mu}_1(0)/\sigma_1(0)}{\hat{\mu}_2(0)/\sigma_2(0)} \right)^2$$

for every sequence $\theta_\nu \downarrow 0$ independently of $\alpha$ and $\gamma$. 

8
Proof. A somewhat tedious but straightforward argument shows that each of $n_i(\alpha, \gamma, \theta_\nu)$, $i = 1, 2$, must diverge to $\infty$. Also to ensure asymptotic size of $\alpha$, the critical value for maybe taken as $\mu_\nu(0) + \sigma_\nu(0)(z_\alpha + r_{n,i})$ where $r_{n,i} \to 0$. And hence,

$$\pi_{n_i}(\alpha, \gamma, \theta_\nu, i)(\theta_\nu) = 1 - \Phi \left( z_\alpha + r_n - \sqrt{n_i(\alpha, \beta, \theta_\nu)\hat{\theta}_\nu \sigma_i(0)}(1 + r^*_n) \right) + \tilde{r}_n$$

where all of $r_n, r^*_n, \tilde{r}_n \to 0$. These power values converge to $\gamma$ only when the argument on $\Phi()$ above converges to $z_\gamma$ and hence,

$$\lim_{\nu \to \infty} \frac{n_2(\alpha, \gamma, \theta_\nu)}{n_1(\alpha, \gamma, \theta_\nu)} = \lim_{\nu \to \infty} \frac{n_2(\alpha, \gamma, \theta_\nu)\theta_\nu^2}{n_1(\alpha, \gamma, \theta_\nu)\theta_\nu^2} = \frac{(z_\alpha - z_\gamma)^2}{(\hat{\mu}_2(0)/\sigma_2(0))^2} = \left( \frac{\hat{\mu}_1(0)/\sigma_1(0)}{\hat{\mu}_2(0)/\sigma_2(0)} \right)^2.$$

\[\square\]

3.2 Le Cam’s Third Lemma & Local Asymptotic Normality

The regularity property (5) requires establishing asymptotic normality of $T_n$ under a sequence of probability distributions $q_n = q_{n, \theta_0 + h/\sqrt{n}}$ associated with a sequence of parameter values $\theta_n = \theta_0 + h/\sqrt{n}$. We have seen that it is often possible to establish asymptotic normality under the sequence $p_n = p_{n, \theta_0}$ associated with the fixed parameter value $\theta_0$. Le Cam developed a brilliant theory of contiguity and local asymptotic normality to show that asymptotic normality under $p_n$ could be enough to establish asymptotic normality under $q_n$. We will not see his theory in full glory, but here is the much celebrated “Le Cam’s third lemma”:

**Theorem 3** (Le Cam’s third lemma). For every $n = 1, 2, \ldots$, let $p_n, q_n$ be two pdfs defined on a common space $X_n$ and let $T_n : X_n \to \mathbb{R}^k$ be a map such that,

$$\left( \begin{array}{c} T_n(X^{(n)}) \\ \log \frac{q_n}{p_n}(X^{(n)}) \end{array} \right) \xrightarrow{\mathcal{L}} N_{k+1}\left( \left( \begin{array}{c} \mu \\ -\frac{1}{2}\Sigma^2 \end{array} \right) \cdot \left( \begin{array}{cc} \tau \\ \sigma^2 \end{array} \right) \right) \quad \text{under } X^{(n)} \sim p_n.$$

Then $T_n(X_n) \xrightarrow{\mathcal{L}} N_k(\mu + \tau, \Sigma)$ under $X^{(n)} \sim q_n$.

The result might look a peculiar at first because of the requirement that $\log \frac{q_n}{p_n}(X^{(n)}) \xrightarrow{\mathcal{L}} N(-\sigma^2/2, \sigma^2)$ for some $\sigma^2$. There is nothing unusual about this because it holds for all regular models. In fact it holds under a broad characterization known as “Local Asymptotic Normality (LAN)”. Here is the definition:

**Definition 2** (Local Asymptotic Normality). A sequence of statistical models $(p_{n, \theta} : \theta \in \Theta)$ is called *locally asymptotically normal (LAN)* at $\theta = \theta_0$ if there exist matrices $r_n, I_{\theta_0}$ and random vectors $\Delta_{n, \theta_0}$ such that

$$\Delta_{n, \theta_0} \xrightarrow{\mathcal{L}} N(0, I_{\theta_0}) \quad \text{under } p_{n, \theta_0}$$

and for every converging sequence $h_n \to h$

$$\log \frac{p_{n, \theta_0 + r_n h_n}}{p_{n, \theta_0}}(X^{(n)}) = h^T\Delta_{n, \theta_0} - \frac{1}{2}h^TI_{\theta_0}h + R_n,$$

with $R_n \to 0$ in probability under $p_{n, \theta_0}$.
Note that if the family is LAN at \( \theta_0 \), then
\[
\log \frac{p_{n,\theta_0 + h_n}}{p_{n,\theta_0}}(X^{(n)}) \overset{d}{\to} N(-\frac{1}{2}\sigma^2, \sigma^2)
\]
under \( p_{n,\theta_0} \) with \( \sigma^2 = h^T I_{\theta_0} h \). Therefore for a model satisfying LAN, we can invoke Le Cam’s third lemma for any statistic \( T_n = T_n(X^{(n)}) \) as soon as we can establish a bivariate CLT between \( T_n \) and \( h^T \Delta_{n,\theta_0} \) and calculate the limiting covariance vector \( \tau \). Before we work out a specific example, here are some results on when LAN happens.

**Example (IID models).** Suppose \( p_{n,\theta} \)'s are associated with an IID observations model \( X_1, X_2, \ldots \overset{iid}{\sim} g(x|\theta), \theta \in \Theta \) as in (4). If \( \{g(\cdot|\theta) : \theta \in \Theta\} \) satisfies the Cramér conditions at a \( \theta_0 \) then \( (p_{n,\theta} : \theta \in \Theta) \) is LAN at \( \theta_0 \) with
- \( I_{\theta_0} = I_F(\theta_0) \), the Fisher information at \( \theta_0 \),
- \( r_n = \sqrt{n} I_d \), (\( I_d \) is the \( d \times d \) identity matrix).
- \( \Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} \log g(X_i|\theta_0) = \frac{1}{\sqrt{n}} \hat{\ell}_n(\theta_0) \).

The same assertion holds under a conditions weaker than the Cramér conditions. The family \( \{g(\cdot|\theta) : \theta \in \Theta\} \) is said to be “differentiable in quadratic mean (DQM)” at \( \theta_0 \) if there is a function \( s_{\theta_0}(x) \) such that
\[
\lim_{h \to 0} \frac{1}{h^2} \int \left[ \sqrt{g(x|\theta_0 + h)} - \sqrt{g(x|\theta_0)} - \frac{1}{2} h^T s_{\theta_0}(x) \frac{\sqrt{g(x|\theta_0)}}{g(x|\theta_0)} \right]^2 dx = 0.
\]

When \( \theta \mapsto \log g(x|\theta) \) is differentiable at \( \theta_0 \) for almost every \( x \), one typically has \( s_{\theta_0}(x) = \nabla_{\theta_0} \log g(x|\theta_0) \), the score function. If \( \{g(\cdot|\theta) : \theta \in \Theta\} \) is DQM at \( \theta_0 \) then the associated sequence of models \( (p_{n,\theta} : \theta \in \Theta) \) is LAN at \( \theta_0 \) with \( I_{\theta_0} = E\{s_{\theta_0}(X)s_{\theta_0}(X)^T|\theta_0\} \), \( r_n = \sqrt{n} I_d \) and \( \Delta_{n,\theta_0} = \sum_{i=1}^n s_{\theta_0}(X_i)/\sqrt{n} \).

### 3.3 Slope calculation for Mann-Whitney test

Under the assumed model on \( X^{(n)} = (W_1, \ldots, W_m, V_1, \ldots, V_k) \), we have,
\[
p_{n,\theta}(w_1, \ldots, w_m, v_1, \ldots, v_k) = \prod_{i=1}^m g(w_i - \theta) \prod_{j=1}^m g(v_j) = p_{m,\theta}^1(w^{(m)}) p_{k}^2(v^{(k)}).
\]

Assume \( g \) is differentiable everywhere with derivative \( \hat{g} \). Then LAN holds at \( \theta = 0 \) for the sequence of models \( p_{m,\theta}^1 \) on \( W^{(m)} = (W_1, \ldots, W_m) \), with \( \Delta_{m,0} = -\frac{1}{\sqrt{m}} \sum_{i=1}^m \hat{g}(W_i)/g(W_i) \) and \( r_m = \sqrt{m} \); and the same is inherited by the original model sequence \( p_{n,\theta} \) with \( r_n = \sqrt{m} \) and
\[
\Delta_{n,0} = -\frac{1}{\sqrt{m}} \sum_{i=1}^m \hat{g}(W_i) g(W_i).
\]

We have earlier seen that the centered and scaled Mann-Whitney statistic \( T_n = \sqrt{n}(U - \frac{1}{2}) \) admits the asymptotic expansion (3) with \( \theta = 0 \). And hence if \( m/n \to \lambda \in (0,1) \), then the
Theorem 4 that attain these bounds. Here is a precise statement.

\[ n \text{ there is non-trivial bound on } \lim_{n \to \infty} n \left( 1 - \frac{1}{2} \right) \text{ for various choices of } g. \]

The assumption of the Le Cam’s third lemma holds for \( p_n = p_{n,0} \) and \( q_n = p_{n,h/\sqrt{n}} = p_{n,h_n/r_n} \) where \( h_n = h\sqrt{n}/n \to \lambda \), with \( \mu = 0, \Sigma = 1/\{12\lambda(1 - \lambda)\} \) and

\[ \tau = -hE \left\{ \frac{\hat{g}(W_1)}{g(W_1)} G(W_1) | \theta = 0 \right\} = -h \int G(w) \hat{g}(w) dw = h \int g(w)^2 dw \]

by the integration by parts identity. And so \( \sqrt{n}(U - 1/2) \Rightarrow N(\tau, 1/\{12\lambda(1 - \lambda)\}) \) under \( q_n \). Consequently,

\[ \sqrt{n}\{U - \mu(h/\sqrt{n})\} = \sqrt{n}(U - 1/2) - \sqrt{n}\{\mu(h/\sqrt{n}) - \mu(0)\} \Rightarrow N(\tau - h\mu(0), \frac{1}{12\lambda(1 - \lambda)}) \]

but, \( \mu(0) = \frac{\partial}{\partial \theta} \mu(\theta) |_{\theta = 0} = \frac{\partial}{\partial \theta} \left[ \int [1 - G(w - \theta)] g(w) dw \right] |_{\theta = 0} = \int g(w)^2 dw \) by interchanging differentiation and integration. Hence, \( \sqrt{n}(U - \mu(h/\sqrt{n})) \Rightarrow N(0, \frac{1}{12\lambda(1 - \lambda)}) \) under \( q_n \). So the regularity condition (5) holds with \( \mu(\theta) = P(V_1 < W_1 | \theta) \) and \( \sigma^2(\theta) = 1/\{12\lambda(1 - \lambda)\} \). So the slope is \( \hat{\mu}(0)/\sigma(0) = \sqrt{12\lambda(1 - \lambda)} \int g(w)^2 dw \).

3.4 Relative Efficiency of Mann-Whitney test against T-test

It is rather straightforward to show that the slope of the two sample t-test\(^5\) is \( \sqrt{\lambda(1 - \lambda)}/\sigma_g \), where \( \sigma^2_g = \text{Var}(W_1 | g) = \int w^2 g(w) dw - \{ \int wg(w) dw \}^2 \). So the Pitman relative efficiency of the Mann-Whitney test against the two sample t-test is \( 12\sigma^2_g \int g(w)^2 dw \). Table 1 lists the efficiency number for various choices of \( g \). It can be shown that this number is never smaller that 108/125 \( \approx 86\% \) but it can be made arbitrarily large by considering a thick tailed \( g \).

<table>
<thead>
<tr>
<th>( g(w) )</th>
<th>Relative Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>( \pi^2/9 = 110% )</td>
</tr>
<tr>
<td>Normal</td>
<td>( 3/\pi = 95% )</td>
</tr>
<tr>
<td>Uniform</td>
<td>100%</td>
</tr>
<tr>
<td>( t(3) )</td>
<td>124%</td>
</tr>
<tr>
<td>( t(5) )</td>
<td>190%</td>
</tr>
<tr>
<td>( \frac{3}{4}(1 - w^2)I(</td>
<td>w</td>
</tr>
</tbody>
</table>

Table 1: Relative efficiency of the Mann-Whitney test against the two sample t-test for various choices of \( g \).

4 Parametric Models, Efficiency Bounds and Optimality of ML tests

It turns out that under LAN conditions (and hence Cramér conditions which imply LAN), there is non-trivial bound on \( \lim_{n \to \infty} n(\theta_0 + h/r_n) \) and there are sequence of test procedures that attain these bounds. Here is a precise statement.

\textit{Theorem 4. Let } \( \Theta \subset \mathbb{R}^d \) \text{ and suppose the sequence of parametric models } (p_{n,\theta} : \theta \in \Theta), \ n = 1, 2 \ldots \text{ is LAN at } \theta = \theta_0 \text{ with associated } r_n I_d, I_{\theta_0} \text{ and } \Delta_{n,\theta_0} \text{ such that } r_n \to \infty. \text{ Suppose}

\(^5\)Holds for either version: equal or unequal variance. See HW 2.
\( \eta : \Theta \to \mathbb{R} \) is differentiable at \( \theta_0 \) with non-zero gradient \( \dot{\eta}(\theta_0) \) and \( \eta(\theta_0) = 0 \). Then the power function sequence \( \pi_n(\theta) \) of any sequence of level \( \alpha \) tests for testing \( H_0 : \eta(\theta) \leq 0 \) vs. \( H_1 : \eta(\theta) > 0 \) satisfy, for every \( h \) such that \( \dot{\eta}(\theta_0)^T h > 0 \),

\[
\limsup_{n \to \infty} \pi_n(\theta_0 + \frac{h}{r_n}) \leq 1 - \Phi \left( z_\alpha - \frac{\dot{\eta}(\theta_0)^T h}{\sqrt{\dot{\eta}(\theta_0)^T I_{\theta_0}^{-1} \dot{\eta}(\theta_0)}} \right)
\]

and the bound is attained by tests “reject \( H_0 \) if \( T_n \geq z_\alpha \)” with

\[
T_n = \frac{\dot{\eta}(\theta_0)^T I_{\theta_0}^{-1} \Delta_{n,\theta_0}}{\sqrt{\dot{\eta}(\theta_0)^T I_{\theta_0}^{-1} \dot{\eta}(\theta_0)}}
\]

A sketch of a proof. It can be established that (perhaps on some subsequence) \( \pi_n(\theta_0 + h/r_n) \) converges to some \( \pi(h) \). The LAN property then leads to the fact that \( \pi(h) \) is the power function for some size \( \alpha \) procedure for testing \( H_0 : \dot{\eta}(\theta_0)^T h \leq 0 \) vs. \( H_1 : \dot{\eta}(\theta_0)^T h > 0 \) in the model \( (N_d(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k) \). But there is a size \( \alpha \) UMP test for this \( H_0 \) in this model, whose power function is precisely the bound displayed in the statement of the theorem. That the test “reject \( H_0 \) if \( T_n \geq z_\alpha \)” attains the bound can shown by establishing \( T_n \overset{d}{\to} N_d(\dot{\eta}(\theta_0)^T h / \{\dot{\eta}(\theta_0)^T I_{\theta_0}^{-1} \dot{\eta}(\theta_0)\}^{1/2}, 1) \) under \( \theta_0 + h/r_n \) by applications of Le Cam’s third lemma and Slutsky’s theorem.

\[\text{\textsuperscript{6}}\text{This requires a powerful “representation theorem” which we omit.}\]