Bios 761 Abridged Notes

1 Introduction

This is a set of abridged notes for BIOS 761 at the University of North Carolina compiled by Ethan Alt. These notes serve to help study for the PhD Qualifying Exam. The goal is to present key ideas and theorems from the course in note form rather than slide form. If there are any errors, please contact Ethan Alt at ethanalt@live.unc.edu.

2 Decision Theory

2.a Basic Elements

The unknown quantity θ is often called the state of nature. The parameter space Θ denotes all possible states of nature. Decisions are more commonly called actions. A particular action will be denoted by 'a' and the set of all possible actions by $\mathcal{A}$ = the action space. $\mathcal{X}$ = the sample space of a random variable X with distribution P_θ. X can be continuous and have a density with respect to Lebesgue measure, or X can be discrete and have density (probability function) with respect to a counting measure.

The key element in decision theory is the loss function. If an action a_1 is taken and θ_1 turns out to be the true state of nature, then a loss $L(\theta_1, a_1)$ will be incurred. The loss function, $L(\theta, a)$, is defined for all $(\theta, a) \in \Theta \times \mathcal{A}$, where $\Theta \times \mathcal{A}$ is the Cartesian product of the sets Θ and $\mathcal{A}$

$$\Theta \times \mathcal{A} = \{(\theta, a) : \theta \in \Theta, a \in \mathcal{A}\}$$

Thus, $L : \Theta \times \mathcal{A} \to \mathcal{R}^+$, since we will not consider negative losses (gaines) in this course.

Definition 1.1 (Non-randomized decision rule)
A non-randomized decision rule, $d(X)$, is a function from $\mathcal{X}$ into the action space $\mathcal{A}$, ($d : \mathcal{X} \to \mathcal{A}$, such that if $X = x$ is observed, we take action $d(x)$ with probability 1 (non-randomized).

Definition 1.2 (Risk Function)

The risk function of a decision rule $d(x)$ is the expectation of the loss function. That is,

$$\begin{aligned} R(\theta, d) &= E_\theta[L(\theta, d(x))] \\ &= \int_{\mathcal{X}} L(\theta, d(x)) P_\theta(dx) \\ &= \int_{\mathcal{X}} L(\theta, d(x)) p(x \mid \theta) \end{aligned}$$

When there is no data (no data problem), $R(\theta, d) = L(\theta, d)$ (there is no probability function to integrate over). In this case, $R(\theta, d)$ is referred to as the frequentist risk.

Definition 1.3 (Inadmissible Rules)
A decision rule d is inadmissible if there exists a rule d' such that

$$R(\theta, d') \leq R(\theta, d)$$

for all $\theta \in \Theta$ with strict equality holding for at least one θ. Note that this definition means that the risk of d' is uniformly smaller than the risk of d. We define a decision rule to be admissible if it is not inadmissible.

Definition 1.4 (Randomized Decision Rule)
A randomized decision rule $d(a, x)$ is, for each x, a probability distribution on $\mathcal{A}$. That is, $d : \mathcal{A} \times \mathcal{X} \to [0, 1]$ and $d(a, x) := d(a|X) =$ probability of action a when $X = x$ is observed.

Definition 1.5 (Loss Function of Randomized Rule)
The loss function $L(\theta, d(\cdot|x))$ of a randomized rule $d(a \mid x)$ is defined to be

$$\begin{aligned} L(\theta, d(\cdot \mid x)) &= E_{d(\cdot|x)}[L(\theta, a)] \\ &= \int L(\theta, a) d(da|X = x) \\ &= \int L(\theta, a) p(a|x) da \end{aligned}$$

where $p(a|x)$ is the probability you take action a given $X = x$ is observed.

Definition 1.6 (Risk Function of Randomized Decision Rule)
The risk function for a randomized rule $d \in \mathcal{D}$ when $\theta \in \Theta$ is defind as

$$\begin{aligned} R(\theta, d) &= E_\theta[L(\theta, d(\cdot|x))] \\ &= \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) d(da|X = x) P_\theta(dx) \end{aligned}$$

Remark 1.1
Note that nonrandomized decision rules are just special cases of randomized decision rules. In particular, if $d(x)$ is a nonrandomized rule, then the equivalent randomized rule is

$$I_a(d(x)) = \begin{cases} 1 & \text{if } d(x) \in a \\ 0 & \text{if } d(x) \notin a \end{cases}$$

Definition 1.7 (Optimal Rules)
A decision rule d is said to be optimal if

1. d is admissible, and
2. for any other admissible rule d', $d' = d$

Thus, optimal rules are unique admissible rules.

Definition (Minimax Principle)
A decision rule d_M is minimax if

$$\inf_{d \in \mathcal{D}} \left\{ \sup_{\theta \in \Theta} R(\theta, d_1) \right\} = \sup_{\theta in \Theta} R(\theta, d_M)$$

That is, a rule is minimax if it minimizes the worst possible risk $\sup_{\theta \in \Theta} R(\theta, d)$ among all randomized rules $d \in \mathcal{D}$.

The concept of minimaxity is best explained within a Bayesian framework. It turns out that Bayes estimators with proper priors are minimax and admissible.

Definition 1.9 (Prior Distribution)
A probability distribution Λ over Θ is called a prior distribution. Note that this distribution is over the parameter space Θ.

Definition 1.10 (Bayes Risk)
For a given prior Λ and $d \in \mathcal{D}$, the Bayes Risk of d with respect to Λ is

$$R(\Lambda, d) = \int_{\Theta} R(\theta, d) d\Lambda(\theta)$$

Definition 1.11 (Bayes Rule)
A Bayesian decision rule with respect to Λ, denoted as d_Λ, is any rule satisfying

$$\mathcal{R}(\Lambda, d_\Lambda) = \inf_{d \in \mathcal{D}} R(\Lambda, d)$$

that is, a Bayes rule is a rule that minimizes Bayes risk among all other rules with respect to some prior Λ.

A more convenient way of finding Bayes rules is to first write

$$R(\theta, d) = \int_{\mathcal{X}} L(\theta, d(x)) p(x|\theta) dx$$

By a change in the order of integration, we can write

$$R(\Lambda, d) = \int_{\mathcal{X}} \left[\int_{\Theta} L(\theta, d(x)) p(\theta|x) d\theta \right] p(x) dx$$

To minimize the double integral, we may minimize the inside integral separately for each x, that is, for each x, we minimize

$$\int L(\theta, d(x))p(\theta|x)d\theta$$

This quantity is called the posterior expected loss. This leads us to our first theorem.

Theorem 1.1 A Bayes Rule d_λ is a rule which minimizes the posterior expected loss (wrt $d(x)$)

$$d_\Lambda = \arg\min_{d(x)} \left[\int L(\theta, d(x))p(\theta|x)d\theta \right]$$

Note that to minimize the Bayes Risk with respect to some action a, we can simply take the derivative of the Posterior Expected Loss with respect to a, which can be moved to the inside of the integral (since the integral is taken over θ).

Theorem 1.2
If $L(\theta, a) = c|\theta - a|$ for some $c > 0$, the Bayes estimator for θ is the posterior median.

Theorem 1.3
If

$$L(\theta, a) = \begin{cases} 0, |\theta - a| \leq c, & c > 0 \\ 1, |\theta - a| > c, & c > 0 \end{cases}$$

then the Bayes estimator of θ converges to the posterior mode of θ as $c \to 0$.

Theorem 1.4
If

$$L(\theta, a) = \begin{cases} k_1, |\theta - a| \leq c, & c > 0 \\ k_2, |\theta - a| > c, & c > 0 \end{cases}$$

for some $k_1 > 0$ and $k_2 > 0$, then the Bayes estimator of θ is the posterior p-th quantile, where p depends on k_1 and k_2.

Definition 1.12 (Convex Set)
$\mathcal{A} \subset \mathcal{R}^l$ is convex if, for all $\lambda \in [0, 1]$, $x, y \in \mathcal{A}$,

$$\lambda x + (1 - \lambda)y \in \mathcal{A}$$

In words, a set is convex if it is closed under convex combinations–linear combinations with λ and $1 - \lambda$ where $\lambda \in [0, 1]$.

Theorem 1.5 (Convexity of Risk Body)
Define the risk body as

$$\mathbf{R} = \{(R(\theta_1, d), \ldots, R(\theta_l, d)) \in \mathcal{R}^{+l} : d \in \mathcal{D}\}$$

Then the risk body is convex.

Theorem 1.6
Every $d \in \mathcal{D}$ may be expressed as a convex combination of the nonrandomized rules.

Theorem 1.7
If the risk set contains its boundary points, then a minimax rule always exists

Theorem 1.9 (Admissibility of Bayes Rules)
Suppose that d_0 is a Bayes rule with respect to $\Lambda = (\lambda_1, \ldots, \lambda_l)$ and $\lambda_i > 0$. Then d_0 is admissible

Theorem 1.10 (Bayes Rules w/ Constant Risk are Minimax)
If $d_\Lambda \in \mathcal{D}$ is Bayes for Λ, and it has constant risk, then d_Λ is minimax.

Definition 1.14 (Conjugate Prior)
Suppose X has density $p(x|\theta)$ and λ is a prior density, $\lambda \in P_\Lambda$ for θ. If $\lambda(\theta|x)$ has the same form as the prior, i.e., $\lambda(\theta|x) \in P_\Lambda$ then λ is said to be a conjugate prior for θ

Below, we present a table on conjugate priors.

$X\|\theta$	$\Lambda(\theta)$
Bernoulli	Beta
Binomial	Beta
Negative Binomial	Beta
Poisson	Gamma
Hypergeometric	Beta-Binomial
Geometric	Beta
Multinomial	Dirichlet
Normal (known variance)	Normal
Normal (known mean)	Inverse Gamma
Gamma	Gamma

Definition (Generalized Bayes Rule)
A rule d_{GB} is said to be a generalized Bayes rule if there exists a measure $\lambda(\theta)$ on Θ such that

$$\int L(\theta, d)p(\theta|x)d\theta$$

takes on a finite minimum when $d = d_{GB}$.

Note: the definition says measure, not probability measure. Thus, generalized Bayes rules are Bayes rules that need not have proper priors.

Theorem 1.11
Suppose $\theta \sim \Lambda$, $(x \mid \theta) \sim P_\theta$, and $L(\theta, a) \geq 0$ for all (θ, a). If

1. there exists a rule d_0 with finite risk, and
2. there exists a rule d_Λ minimizing the posterior expected loss for each x, then $d_\Lambda(x)$ is a Bayes rule.

Corollary 1.1 (Estimation with Weighted Squared Error Loss)
If $\mathcal{A} = \mathbf{\Theta}$ (estimation problem) and $L(\theta, a) = k(\theta)|\theta - a|^2$ (loss function is weighted squared

error loss with weights, then

$$d_\Lambda(x) = \frac{E[k(\theta)]\theta|X = x]}{E[k(\theta)|X = x]}$$

Corollary 1.2 (Estimation with L_1 norm)
If $\mathbf{\Theta} = \mathcal{A}$ and $L(\theta, \mathbf{a}) = |\theta - \mathbf{a}|$, then $d_\Lambda(x) =$ any median of $\Lambda(\theta|x)$.

2.b How to find Minimax Rules

2.b.1 Least Favorable Priors

Definition 1.16 (Least Favorable Prior)
A prior Λ_0 for which $\mathcal{R}(\Lambda, d_\Lambda)$ is <u>maximized</u> is called a <u>least favorable prior</u>:

$$\mathcal{R}(\Lambda_0, d_{\Lambda_0}) = \sup_\Lambda \mathcal{R}(\Lambda, d_\Lambda)$$

Theorem 1.12
Suppose that Λ is a prior distribution on Θ such that

$$\begin{aligned} R(\Lambda, d_\Lambda) &= \int_\Theta R(\theta, d_\Lambda) d\Lambda(\theta) \\ &= \sup_\theta R(\theta, d_\Lambda) \end{aligned}$$

Then

1. d_Λ is minimax
2. If d_Λ is unique Bayes wrt Λ, then d_Λ is unique minimax
3. Λ is least favorable.

Corollary 1.5 (Bayes w/ constant risk are minimax)
If d_Λ is Bayes wrt Λ and has constant risk, then d_Λ is minimax.

2.b.2 Sequence of Least Favorable Priors

Definition 1.17
The sequence of prior distributions $\{\Lambda_k\}$ with Bayes Risks $\{\mathcal{R}_k\}$ is said to be <u>least favorable</u> if $\mathcal{R}_\Lambda := \mathcal{R}(\Lambda, d_\Lambda) \leq r$ for any prior distribution Λ.

Theorem 1.13
Suppose that $\{\Lambda_k\}$ is a sequence of prior distributions with Bayes risks satisfying

1. $\mathcal{R}_k \to r$
2. d is an estimator for which $\sup_{\theta\in\Theta}(R(\theta, d)) = r$

then

1. d is minimax, and
2. $\{\Lambda_k\}$ is least favorable.

Theorem 1.13*
Assume that $\{\Lambda_k\}$ is a sequence of proper priors and d is a decision rule such that

$$R(\theta, d) \leq \lim_{k\to\infty} \mathcal{R}(\Lambda_k, d_{\Lambda_k}) < \infty, \text{ for all } \theta$$

where d_{Λ_k} is the Bayes rule corresponding to $\{\Lambda_k\}$. Then d is a minimax rule.

Theorem 1.13 (and Theorem 1.13*) suggests that a good way to come up with a minimax rule is to guess a least favorable prior distribution, whose limit is an improper prior, i.e. $\lim_{k\to\infty} \Lambda_k = \Lambda_0$ where Λ_0 is improper, and then derive the generalized Bayes rule for Λ_0. Such a generalized Bayes rule will often be minimax via Theorem 1.13.

Lemma 1.3
Suppose that $x \sim P \in \mathbf{M} := \{\text{all } P\text{'s on } \mathcal{X}\}$ and that $\nu : P \subset \mathbf{M} \to \mathcal{R}^1$ is a functional. Suppose that d is a minimax estimator of $\nu(P)$ for $P \in P_0 \subset P_1$. If

$$\sup_{P\in P_0} R(P, d) = \sup_{P \in P_1} R(P, d)$$

then d is minimax for estimating $\nu(P), P \in P_1$.

Theorem 1.14 (Admissibility of Bayes Rules)
Any Bayes estimator with finite Bayes Risk is admissible.

Theorem 1.18
Assume that the risk functions $R(\theta, d)$ are continuous in θ for all decision rules d. Assume also that the prior Λ gives positive probability to any open subset of $\mathbf{\Theta}$. Then if $\mathcal{R}(\Lambda, d_\Lambda) < \infty$, the Bayes rule d_λ with respect to Λ is admissible.

2.c Summary of how to find minimax rules

Below, we present a summary of four ways to find minimax rules

1. Direct method: find d_M such that

$$\inf_{d\in\mathcal{D}} \sup_{\theta\in\Theta} R(\theta, d) = \sup_{\theta\in\Theta} R(\theta, d_M)$$

2. Use Theorem 1.12 to find a d_Λ such that

$$R(\Lambda, d_\Lambda) = \sup_{\theta\in\Theta} R(\theta, d_\Lambda)$$

3. Use Theorem 1.13 and 1.13*. Find a sequence of proper priors $\{\Lambda_k\}$ such that the Bayes risks $r_k \to r$ and find a d such that

$$\sup_{\theta \in \Theta} R(\theta, d) = r$$

   That is, we find a least favorable prior distribution which yields a limiting Bayes risk r and then find a rule d with $\sup_{\theta \in \Theta} R(\theta, d) = r$.

4. If d_Λ is Bayes with constant risk, then d_Λ is minimax (Corollary 1.5). Rules with constant risk are sometimes referred to as equalizer rules.

3 Hypothesis Testing

We have a (parametric) model $p(x|\theta) := p_\theta(x)$, and we wish to test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1, \Theta = \Theta_0 \cup \Theta_1$

We will denote the test procedure by $\phi(X)$ = critical function (decision rule). $\phi(X)$ gives us the form of the rejection region for rejecting (not rejecting) H_0, i.e. $\phi(x)$ denotes the decision to reject H0 when X = x is observed. Here, $X = (X_1; ; X_n)$ denotes a random sample of size n.

3.a Simple vs. Simple Tests and the Neyman-Pearson Lemma

We first consider testing a simple null versus a simple alternative, and thus $\Theta_0 = \{0\} := \{\theta_0\}$ and $\Theta_1 = \{1\} := \{\theta_1\}$.

$$\begin{aligned}\text{Let } \alpha &= \text{ significance level or size}\\ &= E_0[\phi(x)]\\ &= \int \phi(x) p_{\theta_0}(x) d\mu\\ &= \int_{\mathcal{X}} \phi(x) p_{\theta_0}(x) dx\end{aligned}$$

$$\begin{aligned}\beta &= \text{power}\\ &= E_1[\phi(x)]\\ &= \int \phi(x) p_{\theta_1}(x) d\mu\\ &= \int \phi(x) p_{\theta_1}(x) dx\end{aligned}$$

$\alpha = P(\text{type I error}) = P(\text{rejecting H0 when H0 is true})$
$1 - \beta = P(\text{type II error}) = P(\text{accepting H0 when H0 is false})$.

The basic approach is to select the critical function $\phi(x)$ which maximizes power $\beta = E_{\theta_1}(\phi(x))$ subject to the condition

$$E_\theta[\phi(x)] \leq \alpha \text{ for all } \theta \in \Theta_0$$

The following theorem is fundamental to the theory of hypothesis testing.

Theorem 2.1 (Neyman-Pearson Lemma)
Let P_0 and P_1 have densitis p_0 and p_1 with respect to some dominating measure μ ($\mu = P_0 + P_1$ always works). Let $0 \leq \alpha \leq 1$. Then

1. There exists a constant k and a critical function ϕ of the form
$$\phi(x) = \begin{cases} 1 & if p_1(x) > kp_0(x) \\ \gamma & if p_1(x) = kp_0(x) \\ 0 & if p_1(x) < kp_0(x) \end{cases}$$
   such that $E_0[\phi(x)] = \alpha$.
2. This test is a most powerful α level test of P_0 vs P_1.
3. If ϕ^* is a most powerful test of size α, then for $a.e.\mu$,
$$\phi^*(x) = \begin{cases} 1 & p_1(x) > kp_0(x) \\ 0 * p_1(x) < kp_0(x) \end{cases}$$

From the proof of the Neyman-Pearson Lemma, we get that

$$\gamma = \frac{\alpha - P_0(Y > k)}{P(Y = k)}$$

is the appropriate γ for the test, where

$$k = \inf\{c : 1 - F_Y(c) < \alpha\}$$

Corollary 2.1
If $0 < \alpha < 1$ and β is the power of the most powerful level α test, then $\alpha < \beta$ unless $P_0 = P_1$.

3.b Composite Null and Alternative Hypotheses

When testing composite null and alternative hypotheses, it is not clear what a "best" or "most powerful" test means, since several values of θ can be entertained in H_0 as well as in H_1. We need new concepts in this situation.

Definition 2.1 (Monotone Likelihood Ratio or MLR)
IF the family of densities $\{p_\theta : \theta \in [\theta_0, \theta_1] \subset \mathcal{R}\}$ is such that $\dfrac{p_{\theta'}(x)}{p_\theta(x)}$ is nondecreasing in $T(x)$ for each $\theta < \theta'$, then the family is said to have a monotone likelihood ratio (MLR).

Definition 2.2 (Size of a Test)
A test ϕ is of size α if

$$\sup_{\theta \in \Theta_0} E_\theta[\phi(x)] = \alpha$$

Let $C_\alpha := \{\phi : \phi \text{ is of size } \alpha\}$. A test ϕ_0 is uniformly most powerful of size α (UMP of size α)

1. ϕ_0 has size α.

2. $E_\theta[\phi_0(x)] \geq E_\theta[\phi(x)]$ for every $\theta \in \Theta_1$ and for every $\phi \in C_\alpha$.

In words, the UMP size α test is a test with size α that has maximum power among all other tests of size α.

Definition 2.3 (Power Function)
The power function of a test ϕ is defined as

$$\beta(\theta) = E_\theta[\phi(x)]$$

Theorem 2.2 (UMP Tests for Composite Hypotheses)
Suppose X has density p_θ with MLR in $T(x)$. Then

- There exists a UMP level α test of $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ which is of the form

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

  with $E_{\theta_0}[\phi(x)] = \alpha$

- $\beta(\theta) := \mathbb{E}_\theta[\phi(x)]$ is increasing in θ for $\beta < 1$.
- For all θ', this same test is the UMP level $\alpha' := \beta(\theta')$ test of $H_0' : \theta \leq \theta'$ vs. $H_1' : \theta > \theta'$
- For all $\theta \leq \theta_0$, the test of (1) minimizes $\beta(\theta)$ among tests satisfying $\alpha = E_{\theta_0}(\phi)$.

The proof of (1) and (2) is done in the following ways:

- Consider a simple vs. simple test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$. Such a UMP test exists by the Neyman-Pearson Lemma and has that specific form. Since the likelihood has MLR, we can write the test in terms of $T(x)$. Now, note that k depends only on α and θ_0, that is, it does not depend on θ_1. Thus, no matter which $\theta_1 > \theta_0$ we choose, the UMP test is the same. As a result, ϕ is UMP for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$.
- Let $\theta' < \theta''$. By NP lemma, a UMP test say ϕ' exists, and by Corollary 2.1, we have

$$E_{\theta'}[\phi'(x)] = \alpha < \beta = E_{\theta''}[\phi'(x)]$$

  Thus, $\beta(\theta)$ is strictly increasing in θ. This means that ϕ in part (1) is UMP for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$. Thus, $\phi(x)$ is UMP out of the class of all tests for which $E_{\theta_0}[\phi(x)] \leq \alpha$.

- We need to show that $\phi(x)$ is UMP for testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Let $\theta' < \theta''$ and consider testing $H_0' : \theta = \theta'$ vs $H_1'\theta = \theta''$. By the Neyman-Pearson Lemma, a UMP test ϕ' exists and by Corollary 2.1, we have

$$\beta(\phi') = E_{\theta'}[\phi'] = \alpha < \beta = E_{\theta''}[\phi'] = \beta(\phi'').$$

Since θ', θ'' were arbitrary, we have that the power function $\beta(\theta)$ is an increasing function of θ. As a result, the ϕ found in (1) satisfies

$$E_\theta[\phi(x)] \leq \alpha \text{ for all } \theta \leq \theta_0$$

Since ϕ is UMP out of the class which satisfies $E_{\theta_0}[\phi(x)] \leq \alpha$, it is UMP out of the class which satisfies $E_\theta \phi(x) \leq \alpha$ for $\theta \leq \theta_0$ since the latter is simply a subclass of the former.

3.c Consistency of NP Tests

Let P and Q be probability measures, and suppose that p and q are their densities with respect to a common σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$. Recall that the Hellinger distance $H(P,Q)$ between P and Q is given by

$$\begin{aligned} H^2(P,Q) &= \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 \, d\mu \\ &= 1 - \int \sqrt{pq} d\mu \\ &= 1 - \rho(P,Q) \end{aligned}$$

where $\rho(P,Q) = \int \sqrt{pq} d\mu$ is the affinity between P and Q. We have $0 \leq \rho(P,Q) \leq 1$

Proposition 1.1
$H(P,Q) = 0 \iff p = q$ a.e.μ $\iff \rho(P,Q) = 1$ Furthermore, $\rho(P,Q) = 0 \iff \sqrt{p} \perp \sqrt{q}$ in the Hilbert space $\mathcal{L}_2(\mu)$.

Proposition 1.2
Let $X_1, \ldots, X_n$ be i.i.d. P (Q) with joint densities

$$p_n(x) = p(x) = \prod i = 1^n p(x_i)$$

$$q_n(x) = q(x) = \prod i = 1^n q(x_i)$$

then $\rho(P_n, Q_n) = [\rho(P,Q)]^n \to 0$ unless $p = q$ a.e. μ.

Theorem 2.4 (Size and power of NP-type Tests)
For testing p vs. q (defined above), the test

$$\phi_n(x) = \begin{cases} 1, & \text{if } q_n(x) > k_n p_n(x) \\ 0, & \text{if } q_n(x) < k_n p_n(x) \end{cases}$$

with $0 < a_1 \leq k_n \leq a_2 < \infty$ for all $n \geq 1$ is size and power consistent: both probabilities of error, $\alpha := \alpha_n \to 0$ and $1 - \beta_n \to 0$ as $n \to \infty$.

3.d Unbiased Tests

Consider testing

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

where $X \sim P_\theta$ for some $\theta \in \Theta = \Theta_0 \cup \Theta_1$. Let ϕ denote the critical function.

Definition 2.4 (Unbiased Tests and Similar on the Boundary Tests)

1. ϕ is unbiased if $\beta_\phi(\theta) \geq \alpha$ for all $\theta \in \Theta_1$ and $\beta_\phi(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.

2. ϕ is said to be similar on the boundary (SOB) if

$$\beta_\phi(\theta) = \alpha \text{ for all } \theta \in [\operatorname{int}(\Theta_0) \cup \operatorname{int}(\Theta_1)]^C := \Theta_B$$

   where $\operatorname{int}(\Theta_1)$ is the interior of Θ_1 This definition says that the power function $= \alpha$ on the boundary of the parameter space.

Remark 2.2

- If ϕ is UMP size α, then ϕ is unbiased
- If ϕ is unbaised and $\beta_\phi(\theta)$ is continuous for $\theta \in \Theta$, then ϕ is SOB.

Definition 2.5 (Uniformly Most Powerful Unbiased (UMPU) Test)
A uniformly most powerful unbiased UMPU level α test is a test ϕ_0 for which

$$E_\theta[\phi_0] \geq E_\theta[\phi] \text{ for all } \theta \in \Theta_1$$

for all unbiased level α tests ϕ.

Lemma 2.1
If $P = \{P_\theta : \theta \in \Theta\}$ is such that $\beta_\phi(\theta)$ is continuous for all test functions ϕ, and if ϕ_0 is UMP among all SOB tests (UMP SOB) for H_0 vs H_1, and if ϕ_0 is level α for H_0 vs H_1, then ϕ_0 is UMPU for H_0 vs H_1.

Remark 2.3

- For a multiparameter exponential family with densities

$$\frac{\mathrm{d}p_\theta(x)}{\mathrm{d}\mu} = c(\theta) \exp\left\{\sum_j \theta_j T_j(x)\right\}$$

  The power function $\beta_\phi(\theta)$ is continuous in θ for all ϕ (see Lemma 2.1)

- UMP $\implies$ UMPU, UMPU $\not\Longrightarrow$ UMP.

3.d.1 Application to One-Parameter Exponential Families

Theorem 2.5

Consider $p_\theta(x) = c(\theta)\exp\{\theta T(x)\}h(x)$ with respect to a σ-finite measure μ on some subset of R^n. We wish to test

(1) $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$

(2) $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ vs. $H_1 : \theta_1 < \theta < \theta_2$

(3) $H_0 : \theta_1 \leq \theta \leq \theta_2$ vs. $H_1 : \theta < \theta_1$ or $\theta > \theta_2$

(4) $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

(1) The test ϕ_1 with $E_{\theta_0}[\phi(T)] = \alpha$ given by

$$\phi_1(T(x)) = \begin{cases} 1 & \text{if } T(x) > k \\ \gamma & \text{if } T(x) = k \\ 0 & \text{if } T(x) < k \end{cases}$$

is <u>UMP</u> of level α for H_0 vs H_1 in (1).

(2) The test ϕ_2 with $E_{\theta_i}[\phi_2(T)] = \alpha, i = 1, 2$ given by

$$\phi_2(T(x)) = \begin{cases} 1 & \text{if } k_1 < T(x) < k_2 \\ \gamma_i & \text{if } T(x) = k_i, i = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

is <u>UMP</u> of level α for H_0 vs H_1 in (2).

(3) The test ϕ_3 with $E_{\theta_i}(\phi_3(T)) = \alpha, i = 1, 2$ given by

$$\phi_3(T(X)) = \begin{cases} 1 & \text{if } T(x) < k_1 \text{ or } T(x) > k_2 \\ \gamma_i & \text{if } T(x) = k_i, i = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

is <u>UMPU</u> of level α for H_0 vs H_1 in (3). A UMP test for (3) does not exist.

(4) The test ϕ_4 with $E_{\theta_0}[\phi_4(T)] = \alpha$ and $E_{\theta_0}[T\phi_4(t)] = \alpha E_{\theta_0}[T]$, given by

$$\phi_4(T(x)) = \begin{cases} 1, & \text{if } T(x) < k_1 \text{ or } T(x) > k_2 \\ \gamma_i, & \text{if } T(x) = k_i, \quad i = 1, 2 \\ 0, & \text{otherwise} \end{cases}$$

is <u>UMPU</u> for H_0 vs H_1 in (4). Furthermore, if T is symmetrically distributed about a under θ_0, then $E_{\theta_0}[\phi_4(T)] = \alpha$, $k_2 = 2a - k_1$ and $\gamma_1 = \gamma_2$ determine the constants.

3.e UMPU Tests for Families with Nuisance Parameters via Conditioning

Definition 1.6 (Neyman Structure)
Let T be sufficient for $P_B := \{P_\theta : \theta \in \Theta_B\}$ and let $P^T := \{P_\theta^T : \theta \in \Theta_B\}$. A test function ϕ is said to have Neyman Structure with respect to T if

$$E(\phi(X)|T) = \alpha \text{ a.s. } P^T.$$

Note: it is not clear in the notes, but it appears Θ_B is as defined in the SOB tests–the complement of the union of the interior spaces.

Theorem 1.6 (When SOB tests have Neyman Structure)
Let X be a random variable with distribution $P_\theta \in P = \{P_\theta : \theta \in \Theta\}$, and let T be sufficient for $P_B = \{P_\theta : \theta \in \Theta_B\}$ Then all SOB tests have Neyman Structure w.r.t. T if and only if the family of distributions $P^T := \{P_\theta^T : \theta \in \Theta_B\}$ is boundedly complete, that is, if $E_P[h(T)] = 0$ for all $P \in P^T$ with h bounded, then $h = 0$ a.e. $\P^T$.

Remark 1.5
For exponential families, all critical functions ϕ have a continuous power functions. If we can show that T is sufficient for P_B and P^T is boundedly complete, then all unbiased tests are SOB and all SOB tests have Neyman Structure. Thus, if we can find a UMP Neyman Structure test ϕ_0 and we can show ϕ_0 is unbiased, then ϕ_0 is UMPU.

UMP Neyman Structure tests can be easier to find, since they are characterized by having conditional probability of rejection equal to α on each surface $T = t$. But the distribution of each surface is independent of $\theta \in \Theta_B$ because T is sufficient for P^T. Thus, the problem has been reduced to testing a one parameter hypothesis for each fixed value of t, and in many problems wecan easily find the most powerful test of this simple hypothesis.

3.f UMPU Tests for Multiparameter Exponential Families

Consider the exponential family $P = \{p_{\theta,\xi}\}$ given by

$$p_{\theta,\xi} = c(\theta, \xi) \exp\left[\theta u(x) + \sum_{i=1}^{k} \xi_i T_i(x)\right]$$

with respect to a σ-finite measure μ on some subset of R^n, where Θ is convex, has dimension $k+1$, $\Theta = \{\theta, \xi_1, \ldots, \xi_k\}$.

Theorem 1.7 (UMPU Test for Multiparameter Exponential Family) We wish to test

(1) $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$

(2) $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ vs. $H_1 : \theta_1 < \theta < \theta_2$

(3) $H_0 : \theta_1 \leq \theta \leq \theta_2$ vs. $H_1 : \theta < \theta_1$ or $\theta > \theta_2$

(4) $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

The following are the UMPU tests for (1) - (4).

(1)

$$\phi_1(x) = \begin{cases} 1 & \text{if } U > c(t) \\ \gamma(t) & \text{if } U = c(t) \\ 0 & \text{if } U < c(t) \end{cases}$$

where $E_{\theta_0}[\phi_1(U)|T = t] = \alpha$

(2)

$$\phi_2(x) = \begin{cases} 1 & \text{if } c_1(t) < U < c_2(t) \\ \gamma_i & \text{if } U = c_i(t) \\ 0 & \text{otherwise} \end{cases}$$

where $E_{\theta_i}[\phi_2(U)|T = t] = \alpha, i = 1, 2$

(3)

$$\phi_3(x) = \begin{cases} 1 & \text{if } U < c_1(t) \text{ or } U > c_2(t) \\ \gamma_i & \text{if } U = c_i(t) \\ 0 & \text{otherwise} \end{cases}$$

where $E_{\theta_i}[\phi_3(U)|T = t] = \alpha, i = 1, 2.$

(4)

$$\phi_4(x) = \begin{cases} 1 & \text{if } c_1(t) < U < c_2(t) \\ \gamma_i & \text{if } U = c_i(t) \\ 0 & \text{otherwise} \end{cases}$$

where $E_{\theta_0}[\phi_4(U)|T = t] = \alpha$ and
$E_{\theta_0}[\phi_4(U)|T = t] = \alpha E_{\theta_0}[U|T = t]$

The remark below gives UMPU tests if an ancillary statistic $h(U,T)$ is found **Remark 2.6** (UMPU Tests for Ancillary Statistics)

(a) If $V = h(U,T)$ is increasing in U for each fixed t and is independent of T on Θ_B then

$$\phi_1(x) = \begin{cases} 1 & \text{if } V > c \\ \gamma & \text{if } V = c \\ 0 & \text{if } V < c \end{cases}$$

is UMPU in hypothesis (1)

(b) If $V := h(U,T) = a(t)U + b(t)$ with $a(t) > 0$, then the second constraint in (4) becomes

$$E_{\theta_0}\left[\frac{V-b(t)}{a(t)}\phi\middle| T=t\right] = \alpha E_{\theta_0}\left(\frac{V-b(t)}{a_t}\phi\middle| T=t\right)$$

or $E_{\theta_0}[V\phi|T=t] = \alpha E_{\theta_0}[V|T=t]$ and if V is independent of T on the boundary, then the test for (4) is unconditional.

(c) For hypotheses (2) and (3), if V is monotone in U and V is independent of T on the boundary, then the test is unconditional.

3.g Likelihood Ratio Tests (LRT)

UMP and UMPU tests may not always exist. For example, in the presence of nuisance parameters, UMPU tests are quite hard to construct, especially when the dimension of the nuisance parameter is high. In k-parameter exponential families, for example, if we want to test one parameter treating all others as nuisance, UMPU tests typically do not exist.

Suppose we are interested in testing $H_0\theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where $\Theta \subset R^k$ so that Θ_0 and Θ_1 are subsets of R^k. Suppose $X_1, \ldots, X_n$ are i.i.d. from $p(x|\theta)$. The likelihood ratio test is defined by

$$\Lambda = \frac{\sup_{\theta\in\Theta_0} p(x \mid \theta)}{\sup_{\theta\in\Theta_0\cup\Theta_1} p(x|\theta)}$$

where $p(x|\theta) = \pi_{i=1}^n p(x_i|\theta)$.

We reject H_0 if Λ is too small. Thus, the likelihood ratio test of level α is given by

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

where k is chosen so that $\alpha = \sup_{\theta\in\Theta_0} E[\phi(x)]$. In some situations, such a k may not exist.

Theorem 2.8 (LRT and UMP/UMPU Tests for Exponential Families) Suppose X has a distribution in the 1 parameter exponential family,

$$p_\theta(x) = c(\theta)\exp\{Q(\theta)T(x)\}h(x)$$

where $Q(\theta)$ is a strictly increasing function of θ.

(i) For testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, there exists a LRT whose rejection region is the same as the UMP test given by Theorem 2.5 part (1).

(ii) For testing $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ vs $H_1 : \theta_1 < \theta < \theta_2$, there exists a LRT whose rejection region is the same as the UMP test given by Theorem 2.5 part (2).

(iii) For testing the hypotheses (3) or (4) of Theorem 2.5, there exists LRT whose rejection region is the same as the UMPU test given by Theorem 2.5, parts (3) and (4) respectively.

3.g.1 Asymptotic Distribution of the Likelihood Ratio Statistic

As indicated in Theorem 2.8 and other examples, the LRT is often equivalent to a test based on a statistic whose distribution under H_0 can be used to determine the rejection region of the LRT with size α. When this technique fails, it is difficult or even impossible to find a LRT with size α, even if the cdf of Λ is continuous. The following theorem shows that in the iid case, we can obtain the asymptotic distribution (under H_0) of Λ so that an LRT having asymptotic significance level α can be obtained.

Theorem 2.9 (Asymptotic Distribution of LRT)
Assume the "usual" regularity conditions for a likelihood function, ($p(x)$ is twice continuously differentiable, $I(\theta)$ exists and is finite). Then,

$$-2 \log \Lambda \overset{H_0}{\to}_d \chi^2_r \text{ as } n \to \infty$$

where $r = \dim(\Theta) - \dim(\Theta_0)$.

There are two tests that are asymptotically equivalent to the LRT. These are the Wald and Score tests.

3.g.2 Wald Test

Suppose we wish to test $H_0 : \theta = \theta_0$. The Wald test rejects H_0 when the value of

$$W_n = (\hat{\theta} - \theta)^T I_n(\hat{\theta})(\hat{\theta} - \theta_0)$$

is large, where $\theta = (\theta_1, \ldots, \theta_r$, $I_n(\theta)$ is the Fisher information based on $(X_1, \ldots, X_n)$ where $\hat{\theta}$ is the MLE of θ.

Theorem 2.10 (Score Test)
Assume the same regularity conditions as Theorem 1.9. Then as $n \to \infty$, $W_n \overset{H_0}{\to} \chi^2_r$, and therefore the test rejects H_0 at a level α if $W_n > \chi^2_{r,1-\alpha}$

3.g.3 Score Test

Rao (1947) introduced a Score test, that rejects $H_0 : \theta = \theta_0$ when the value of

$$R_n = [\nabla \ell(\theta_0)]'[I_n(\theta_0)]^{-1}[\nabla \ell(\theta_0)]$$

is large, where $\nabla \ell(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta(x_i)|_{\theta=\theta_0}$ and $I_n(\theta)$ is the Fisher information.

As $n \to \infty$, $R_n \overset{H_0}{\to} \chi^2_r$.

3.h Bayesian Hypothesis Testing

In classical hypothesis testing, a null hypothesis $H_0 : \theta \in \Theta_0$ and an alternative hypothesis, $H_1 : \theta \in \Theta_1$ are specified. A test procedure is evaluated in terms of the probabilities of Type I and Type II error. These probabilities of error represent the chance that a sample is observed for which the test procedure will result in the wrong hypothesis being accepted. In Bayesian analysis, the task of deciding between H_0 and H_1 is conceptually more straightforward. One merely calculates the posterior probabilities $p(\theta_0|x)$ and $p(\theta_1|x)$ and decides between H_0 and H_1 accordingly. The conceptual advantage is that $p(\Theta_0|x)$ and $p(\Theta_1|x)$ are the actual probabilities of the hypotheses in light of the data and the prior distributions.

The quantity used to test hypotheses in the Bayesian framework is called the Bayes factor. The Bayes factor is the Bayesian analogue of the likelihood ratio test. Suppose we wish to test H_0 versus H_1, and let $\lambda_0 = p(H_0)$ $\lambda_1 = p(H_1)$ denote the prior probabilities of H_0 and H_1 respectively. Let $p(H_0|x)$ and $p(H_1|x)$ denote the posterior probabilties of the hypotheses. Then the Bayes factor in favor of H_0 is defined as the posterior to prior odds of H_0 divided by the posterior to prior odds of H_1. That is,

$$B = \frac{\frac{p(H_0|x)}{p(H_0)}}{\frac{p(H_1|x)}{p(H_1)}}$$

$$= \frac{\frac{p(H_0|x)}{p(H_1|x)}}{\frac{p(H_0)}{p(H_1)}}$$

$$= \frac{p(x|H_0)}{p(x|H_1)}$$

$$= \frac{\int_{\theta_0} p(x|\theta, H_0)\lambda(\theta|H_0)d\theta}{\int_{\theta_1} p(x|\theta, H_1)\lambda(\theta|H_1)d\theta}$$

where $\lambda(\theta|H_i)$ is the prior distribution of θ under $H_i, i = 0, 1$. One immediate feature of the Bayes factor is that $p(x|H_0)$ and $p(x|H_1)$ are obtained by integrating over the parameter space instead of maximizing it.

Remark B is defined only when proper priors are used for θ, and B may be sensitive to the choice of prior distribution and/or the choice of prior hyperparameters.

3.i Efficiency of Tests

Definition 2.7 (Pitman Efficiency)
Pitman efficiency is defined to be the limiting ratio of the sample sizes that produces equal asymptotic power against the same sequence of alternatives.

Suppose that $X_1, \ldots, X_N$ have a joint distribution P_θ where θ is a real-valued parameter. We wish to test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that the T_1 test and the T_2 test are both consistent tests of H_0 versus H_1; and that the T_i test rejects H_0 if the statistic $T_{N,i}$ exceeds the upper α percent point of its distribution when $\theta = \theta_0$. Since both tests are consistent, it's useless to compare their limiting power against fixed alternatives; hence we will compare their power on a sequence of alternatives that approach 0 from above at the rate $1/\sqrt{N}$.

Suppose that for each $c > 0$, the statistics $T_{N,i}$ satisfy

$$P_{\theta_0 + c_N/\sqrt{N}}(T_{N,i} \leq x) \to P(N(c\mu_i, \sigma_i^2) \leq \sigma_i x) = P(N(c\mu_i/\sigma_i, 1) \leq x)$$

for all x as $N \to \infty$ for any sequences of c_N's converging to c. Let the T_i test use N_i observations against the sequence of alternatives $c_{N_i}/\sqrt{N_i}$, where $c_{N_i} \to c_i$.

Equal asymptotic power requires

$$\frac{c_1\mu_1}{\sigma_1} = \frac{c_2\mu_2}{\sigma_2}$$

Equal alternative requires

$$\frac{c_{N_1}}{\sqrt{N_1}} = \frac{c_{N_2}}{\sqrt{N_2}}$$

Solving these simultaneously leads to

$$\frac{N_2}{N_1} = \frac{c_{N_2}^2}{c_{N_1}^2} \to \frac{(\mu_1/\sigma_1^2)^2}{(\mu_2/\sigma_2)^2} = e_{1,2}$$

Note that the efficiency $e_{1,2}$ is independent of the common level of significance level α of the tests, of the particular value of the asymptotic power β, and of the particular sequences that converge to the values of c_1 and c_2 that are specified by the choice of β.

3.j Confidence Regions

Definition: (Uniformly Most Accurate Interval)
A function $\theta_L(x)$ for which

$$P_\theta\{\theta_L(x) \leq \theta'\}$$

is minimized for all $\theta' < \theta$ is a uniformly most accurate lower confidence bound for θ at confidence level 1 - α.

Definition (Confidence Set)
A family of subsets $S(x)$ of the parameter Θ is said to constitute a family of confidence sets at confidence level $1 - \alpha$ if

$$P_\theta\{\theta \in S(x)\} \geq 1 - \alpha \text{ for all } \theta \in \Theta$$

that is, if the random set $S(x)$ covers the true parameter point with probability at least $1 - \alpha$.

Theorem 2.14 (Construction of UMA Confidence Sets)

(i.) For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test for testing $H_0 : \theta = \theta_0$ and for each sample point X, let $S(x)$ denote the set of parameter values

$$S(x) = \{\theta : X \in A(\theta), \theta \in \Theta\}$$

then S(x) is a family of confidence sets for θ at confidence level $1 - \alpha$.

(ii.) If for all θ_0, $A(\theta_0)$ is UMP for testing H_0 at level α against the alternatives H_1, then for each $\theta_0 \in \Theta$, S(x) minimizes the probability

$$P_\theta\{\theta_0 \in S(x)\} \text{ for all } \theta \in \Theta_1$$

among all level $1 - \alpha$ families of confidence sets for θ.

A confidence set can therefore be viewed as a combined statement regarding the tests of the various hypotheses H_0, which exhibits the values for which the hypothesis is accepted $\{\theta \in S(x)\}$ and those for which it is rejected $\{\theta \in S(x)^c\}$.

Theorem 2.15
Let the family of densities $p_\theta(x), \theta \in \Theta$, have the MLR property in $T(x)$, and suppose that the cdf $F_\theta(t)$ of $T = T(x)$ is a continuous function in each of the variables t and θ when the other is fixed.

(i.) There exists a uniformly most accurate confidence bound $\theta_L(x)$ (or $\theta_U(x)$) for θ at each confidence level $1 - \alpha$.

(ii.) If $x = (X_1, \ldots, X_n)$ and $t = T(x)$, and if th equation

$$F_\theta(t) = 1 - \alpha$$

has a solution $\theta = \hat{\theta}$ in Θ, then this solution is unique and $\Theta_L(x) = \hat{\theta}$.

Definition: (Confidence Interval)
A confidence interval for θ at a confidence level $1 - \alpha$ is defined as a set of random intervals with endpoints $(\theta_L(x), \theta_U(x))$ such that

$$P_\theta\{\theta_L(x) \leq \theta \leq \theta_U(x)\} \geq 1 - \alpha \text{ for all } \theta \in \Theta$$

Definition: (Pivotal Quantity)
A function of the data and the parameters whose distribution does not depend on any parameters is called a pivotal quantity (or pivotal). We use pivotals to construct confidence bnounds and confidence intervals.

Lemma
When the statistic $T(x)$, which is used in the construction of a confidence bound, has a continuous distribution with cdf $F_t(\theta)$, the quantity $F_T(T|\theta)$ is pivotal since $F_T(T|\theta) \sim U(0,1)$, independent of θ.

3.j.1 Confidence Intervals in the Presence of Nuisance Parameters

When nuisance parameters ξ are present, the defining condition for a lower confidence bound $\theta_L(x)$ becomes

$$P_{\theta,\xi}(\theta_L(x) \leq \theta) \geq 1 - \alpha \text{ for all } (\theta, \xi)$$

Similarly, confidence intervals for θ at confidence lvel $1 - \alpha$ are defined as a set of random intervals with endpoints $(\theta_L(x), \theta_U(x))$ such that

$$P_{\theta,\xi}(\theta_L(x) \leq \theta \leq \theta_U(x)) \geq 1 - \alpha \text{ for all } (\theta, \xi).$$

In the presence of nuisance parameters, we can still use the ideas established earlier in which we constructed confidence bounds (intervals) by finding the corresponding acceptance region of the UMP test. If we are testing $H_0 : \theta = \theta_0$ and $S(x) = \{\theta : X \in A(\theta_0)\}$ then

$$\theta \in S(x) \iff X \in A(\theta)$$

and hence

$$P_{\theta,\xi}\{\theta \in S(x)\} \geq 1 - \alpha \text{ for all } (\theta, \xi)$$

4 Resampling Methods

4.a The Jackknife

Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator of an unknown parameter θ. The bias of T_n is

$$\text{Bias}(T_n) = \mathbb{E}(T_n) - \theta$$

Let $T_{n-1,i} = T_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ be the estimator of θ based on all the observations excluding X_i. Then the jackknife bias estimator is given by

$$b_{JACK} = (n-1)(\bar{T}_n - T_n)$$

This leads to the bias-reduced jackknife estimator:

$$T_{JACK} = T_n - b_{JACK} = nT_n - (n-1)\bar{T}_n$$

Variance of Jackknife Estimator

$$\begin{aligned} v_{JACK} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{T}_{n,i} - \frac{1}{n} \sum_{i=1}^{n} \tilde{T}_{n,i} \right)^2 \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{(n-1),i} - \frac{1}{n} T_{n-1,i} \right)^2 \end{aligned}$$

where $\tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i}$ are known as the *pseudovalues*.

4.b The Bootstrap

Bootstrap samples with replacement from your data, which is stronger than the Jackknife because Jackknife is simply a special case of bootstrap.

Bootstrap Algorithm Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample and suppose we are interested in $\hat{\theta} = s(\boldsymbol{X})$ for some functions $s : \mathbb{R}^n \to \mathbb{R}$. We can write the bootstrap algorithm for estimation of standard errors as:

1. Select B independent bootstrap samples $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_B^*$
2. Obtain the estimate of θ for each bootstrap sample:

$$\hat{\theta}_b^* = s(X_b^*), \quad b = 1, \ldots, B$$

3. Estimate the standard error of $\hat{\theta}$ by using the sample standard deviation of the B bootstrap replications.

$$\widehat{se(\hat{\theta})} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left\{ \hat{\theta}_b^* - \bar{\hat{\theta}}^* \right\}^2}$$

Bias Estimation The bias of an estimator is calculated as $E_F(\hat{\theta}) - \theta$, where F is the CDF. The bootstrap estimate of bias replaces F with F_n–the empirical CDF used for resampling. In the bootstrap ”world”, the ”truth” is $\hat{\theta}$ and $\hat{\theta}_b$ estimates $\hat{\theta}$, $b = 1, \dots B$. That is, population information is replaced by sample information; sample information is replaced by resample information.

$$\widehat{\text{Bias}(\hat{\theta})}_{boot} = E_{F_n}(\hat{\theta}^*) - \hat{\theta}$$

The bias-corrected estimate of $\hat{\theta}$ is

$$\hat{\theta}_{\text{boot, BC}} = \hat{\theta} - \widehat{\text{Bias}(\hat{\theta})}_{\text{boot}}$$

4.c The Parametric Bootstrap

The bootstrap strategy that we have dealt with so far essentially uses the idea:

Replace F by F_n and use samples from F_n to make inference about θ. This method often goes by the fuller title nonparametric bootstrap because the true distribution F is replaced by a nonparametric estimator F_n.

However, it is often the case that a reasonable parametric model exists for F and this information can be incorporated into the bootstrap idea.

Suppose $\boldsymbol{X} = (X_1, \dots, X_n) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The parametric bootstrap algorithm is as follows:

1. Select B independent bootstrap samples $\boldsymbol{X}_1^*, \dots, \boldsymbol{X}_B^*$ from the $N_p(\hat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ distribution, where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the usual estimates.
2. Obtain an estimate of θ for each bootstrap sample.

$$\hat{\theta}_b = s(\boldsymbol{X}_b^*)$$

3. Estimate the standard error of $\hat{\theta}$ by using the sample standard deviation of the B bootstrap replications:

$$\widehat{se(\hat{\theta})} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \overline{\widehat{\theta^*}} \right)^2}$$

4.d Bootstrap Confidence Intervals

4.d.1 Percentile Method Bootstrap Confidence Interval

The most obvious approach to getting 90% confidence intervals for is to find the upper and lower 5% percentiles of the distribution of the bootstrap replicates $\hat{\theta}^*_b$ b and call those a 90% confidence interval for . This is known as the percentile method and essentially amounts to:

Percentile method for bootstrap confidence intervals

1. Select B independent bootstrap samples $\boldsymbol{X}^*_1, \dots \boldsymbol{X}^*_B$.
2. Obtain the estimate of θ for each bootstrap sample: $\hat{\theta}^*_b = s(\boldsymbol{X}^*_b), \quad b = 1, \dots B$.
3. Order the $\hat{\theta}^*_b$'s:
$$\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)}, \dots, \leq \hat{\theta}^*_{(B)}$$
4. The $100(1-\alpha)\%$ confidence interval for θ is
$$\left(\hat{\theta}^*_{[B\alpha/2]}, \hat{\theta}^*_{[(1-\alpha/2)B]}\right)$$
where $[x]$ represents the largest integer less than or equal to x.

4.d.2 Percentile Method (type-2) Bootstrap Confidence Interval

The algorithm for the type-2 method actually finds a confidence interval for the bias of $\hat{\theta}$, and then simply takes the difference between $\hat{\theta}$ and the interval of the bias.

It reduces the error because the former method has to rely on two approximations: $\hat{\theta} \xrightarrow{p} \theta$, and that of replacing the distribution of $\hat{\theta}$ by $\hat{\theta}^*$. With this method, we work with $\hat{\theta} - \theta$ directly and approximate the distribution with that of $\hat{\theta}^* - \hat{\theta}$. Each of these converge to 0, so the interval should be more accurate.

1. Select B independent bootstrap samples: $\boldsymbol{X}^*_1, \dots, \boldsymbol{X}^*_B$
2. Obtain the estimate of θ for each bootstrap sample, $\hat{\theta}^*_b = s(\boldsymbol{X}^*_b)$, and then compute
$$\hat{\phi}^*_b = \hat{\theta}^*_b - \hat{\theta}$$
3. Order the $\hat{\phi}^*_b$ values:
$$\hat{\phi}^*_{(1)} \leq \hat{\phi}^*_{(2)} \leq \cdots \leq \hat{\phi}^*_{(B)}$$
4. The $100(1-\alpha)\%$ confidence interval for θ is
$$\left(\hat{\theta} - \phi^*_{[B(1-\alpha/2)]},\ \hat{\theta} - \phi^*_{[B(\alpha/2)]}\right)$$

4.d.3 Bootstrap t-method for confidence intervals

This type of confidence interval doesn't rely on *any* approximations. The idea is to get bootstrap resamples of T statistics, which we can approximate with the t distribution.

The algorithm is as follows:

1. Select B independent bootstrap samples: $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_B^*$
2. Obtain the estimate of θ for each bootstrap sample and its <u>bootstrap</u> standard error estimate:
$$\hat{\theta}_b^* = s(\boldsymbol{X}_b^*), \quad b = 1, \ldots, B$$
and $\text{SE}(\hat{\theta}_b^*) = \text{SE}(s(\boldsymbol{X}_b^*))$ Then compute
$$\hat{T}_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\text{SE}(\hat{\theta}_b^*)}$$
3. Order the $\hat{T}_b^*$ values:
$$\hat{T}_1^* \leq \hat{T}_2^* \leq \cdots \leq \hat{T}_B^*$$

There is a subtle issue here that is, perhaps, not apparent on first reading: the requirement of a formula for $\text{SE}(\hat{\theta})$. For $\bar{X}$, this problem is easy. For a general statistic, we don't always have a good estimate for the standard error.

Through the use of the bootstrap, we can find an estimate of $\text{SE}(\hat{\theta})$, as the previous section was devoted to. But we need to find an estimate for the standard error of $\text{SE}(\hat{\theta}_b^*$ for each b. This requires that we do two nested bootstrap.

4.e Theory of the Bootstrap

4.e.1 Consistency of the Bootstrap Distribution

Suppose we are interested in estimating the bootstrap distribution of a statistic, $\hat{T}_n$, based on observations $X_1, \ldots X_n$ from some true joint density, F. That is, we want to estimate the sampling distribution of our statistic:

$$H_F(x) = \Pr\Big(\hat{T}_n < x | F\Big)$$

based on the bootstrap distribution,

$$H_{BOOT}(x) = H_{\hat{F}_n}(x) = \Pr\Big(\hat{T}_n^* < x | \hat{F}_n\Big)$$

where $\hat{T}_n^*$ is an observation generated from $\hat{F}_n$. What does it mean to say that H_{BOOT} is a consistent estimator of H_F.

Definition 3.1 Let ρ be a metric on $\mathcal{F} = \{alldistributionson\mathbb{R}\}$. Then H_{BOOT} is ρ-consistent (weakly ρ-consistent) for H_F if $\rho(H_{BOOT}; H_F) \overset{a.s.}{\to} 0$ as $n \to \infty$.

The most common metric is called *Mallow's Distance*: For two distributions H and G in $\mathcal{F}_{r,s}$ (space for all s-dimensional distributions with finite r-th moment),

$$\tilde{\rho}_r(H,G) = \inf_{\gamma_{X,Y}} (\mathbb{E}||X-Y||^r)^{1/r}$$

where $\gamma_{X,Y}$ is the collection of all possible joint distributions of the pairs (X,Y) whose marginal distributions are H and G. Note that if U and V are random variables from H and G respectively, then

$$\tilde{\rho}_r(U,V) = \tilde{\rho}_r(H,G)$$

Bickel and Freedman (Annals of Statistics, 1981) prove some nice results about Mallow's distance, and use this formulation to establish consistency of H_{BOOT}. Some of their results include:

1. For $X_1, \ldots X_n$ iid $F \in \mathcal{F}_{r,s}$,
$$\tilde{\rho}_r(\hat{F}_n, F) \overset{a.s.}{\to} 0$$

2. $$\tilde{\rho}_r(aU, aV) = |a|\tilde{\rho}_r(U,V)$$

3. $$[\tilde{\rho}_2(U,V)]^2 = [\tilde{\rho}(U-\mathbb{E}U, V-\mathbb{E}V)]^2 + ||\mathbb{E}U\mathbb{E}V||^2$$

4. For $\{U_j\}$ and $\{V_j\}$ sequences of independent observations and $\mathbb{E}U_j = \mathbb{E}V_j$, then
$$\left[\tilde{\rho}_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)\right]^2 \leq \sum_{j=1}^m [\tilde{\rho}_2(U_j, V_j)]^2$$

Below, we present an example of showing consistency for a CLT argument.

Example 1.12

Suppose $X_1, \ldots, X_n$ are iid $F \in \mathcal{F}_{r,s}$ and consider $T_n = \sqrt{n}(\bar{X}_n - \mu)$.

$$\begin{aligned}
\tilde{\rho}_2(H_{BOOT}, H_F) &= \tilde{\rho}_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) \\
&= n^{-1/2}\tilde{\rho}_2\left(\sum_{i=1}^{n}(X_i^* - \bar{X}_n), \sum_{i=1}^{n}(X_i - \mu)\right) && \text{(by (2))} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\tilde{\rho}_2(X_i^* - \bar{X}_n), (X_i - \mu)]^2} && \text{(by (4))} \\
&= \tilde{\rho}_2(X_1^* - \bar{X}_n, X_1 - \mu) && \text{(by exchangeability)} \\
&= \sqrt{[\tilde{\rho}_2(X_1^*, X_1)]^2 - ||\mathbb{E}X_1 - \mathbb{E}X_1^*||^2} && \text{(by (3))} \\
&= \sqrt{\left[\tilde{\rho}_2(\hat{F}_n, F)\right]^2 - ||\mu - \bar{X}_n||^2} && \text{(by definition)} \\
&= o(1), a.s.. && \text{(by (1) and since } \bar{X} \overset{a.s.}{\to} \mu)
\end{aligned}$$

Approach can also be used to establish consistency results for differentiable functions of linear statistics, since these can be linearized using a Taylor series expansion. (See Shao and Tu, Section 3.1.5).

4.f Cross Validation

K-fold cross validation algorithm

1. Split the data into K roughly equal parts.
2. For the k-th part, fit the model to the other $K-1$ parts of the data and calculate the prediction error of thee fitted model when predicting the k-th part of the data.
3. Repeat for $k = 1, 2, \ldots, K$, and combine the K estimates of prediction error.

4.g Bootstrapped Hypothesis Testing

Suppose we want to test the null hypothesis $H_0 : F = G$ and suppose that we have identified a suitable test statistic, $T(X)$, which is a function of the observed data X which comprises n observations from population F and m observations from population G. The test statistic could correspond to an estimate of some parameter that characterizes the difference between the two populations or it could be a score test, a rank test, or anything.

To perform inference, we need to find an *Achieved Significance Level*:

$$ASL = P_{H_0}\left\{T(X^*) > T(X)\right\}$$

where $T(X)$ is the observed test statistic and $T(X^*)$ is the test statistic applied to the random variable X^* which has a distribution specified by the null hypothesis. Our usual bootstrap logic applies and suggests the following algorithm:

Bootstrap test for $H_0 : F = G$

1. Select B independent bootstrap samples of size $n + m$ with replacement from X. Arbitrarily call the first n observations $\boldsymbol{z}$ and the second m observations $\boldsymbol{y}$.
2. Evaluate $T(\boldsymbol{x}_b^*) = \bar{\boldsymbol{z}}^* - \bar{\boldsymbol{y}}^*$
3. Approximate ASL_{BOOT} by

$$\widehat{ASL}_{BOOT} = \# \left\{ T(\boldsymbol{x}_b^*) \geq T(\boldsymbol{x}) \right\} / B$$

   where $T(\boldsymbol{x})$ is the observed test statistic.

5 Topics in High Dimensional Data Analysis

5.a Coordinate Descent Algorithm

Let $R_\lambda(\beta_0, \boldsymbol{\beta}) = \dfrac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2$ and $p(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$. We solve

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{(\beta_0, \boldsymbol{\beta})} R_\lambda(\beta_0, \boldsymbol{\beta}) + \lambda p(\beta)$$

Suppose at the current step, we have the solution $(\tilde{\beta}_0, \ldots, \tilde{\beta}_p)^T$. Now, we keep all other coordinates of $\tilde{\boldsymbol{\beta}}$ and update $\tilde{\beta}_j$ by solving:

$$\begin{aligned} \tilde{\beta}_j^{\text{new}} &= \arg\min_{\beta_j} R_\lambda(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_{\boldsymbol{-j}}, \beta_j) + \lambda \sum_{l \neq j} |\tilde{\beta}_l| + \lambda |\beta_j| \\ &= \arg\min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \tilde{y}_i^{(j)} - x_{ij}\beta_j)^2 \right) + \lambda |\beta_j| \end{aligned} \tag{5.1}$$

where $\tilde{y}_i^{(}j) = \tilde{\beta}_0 + \sum_{l \neq j} x_{il}\beta_l$ is the predicted value of y_i excluding the contribution from x_{ij}. It can be shown that the solution to (5.1) is given by

$$\tilde{\beta}_j^{\text{new}} = S\left(\frac{1}{n} \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}, \lambda \right)$$

where $S(z, \lambda)$ is the soft-thresholding function defined by

$$S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+$$

then we have an important observation:

$$\tilde{\beta}_j^{\text{new}} = S\left(\tilde{\beta}_j - \frac{\partial R_\lambda(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_{-j}, \beta_j)}{\partial \beta_j}, \lambda \right)$$

This means that we first move along the direction of negative gradient with a step-size equal to 1, then we apply the soft-thresholding operation. This is where the name of "coordinate descent" comes from. The soft-thresholding step is due to the L1-penalty.

5.b Subgradients and Subdifferentials for Convex Optimization Problems

One of the reasons this algorithm works is because the lasso problem is a convex optimization problem. Recall that f is a convex function if for any x_1, x_2 and any $t \in [0, 1]$:

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Recall that if a convex function f is differentiable,

$$f(y) \geq f(x) + [\nabla f(x)]^T (y - x)$$

For non-differentiable f, we turn to the concept of subgradients.

Definition: (Subgradient)
g is called a subgradient of f (not necessarily convex) at x if and only if

$$f(y) \geq f(x) + g^T (y - x) \text{ for all } y$$

Definition (Subdifferential)
The set of all subgradients of f at x is called the subdifferential of f at x, and is denoted by $\partial f(x)$.

Subdifferentials are important because they characterize the minimizer of f. In particular, if f is convex and differentiable,

$$f(x^*) = \inf_x f(x) \text{ if and only if } \partial f(x^*) = 0$$

For non-differentiable convex f, we have

$$f(x^*) = \inf_x f(x) \text{ if and only if } 0 \in \partial f(x^*)$$

5.b.1 KKT Conditions

Any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying the following conditions is a Lasso estimator:

$$(\boldsymbol{X^T X \beta})_j - \boldsymbol{X_j^T y} + \lambda_n \text{sign}(\beta_j) = 0 \text{ for } \beta_j \neq 0 \tag{5.2}$$

$$|(\boldsymbol{X^T X \beta})_j - \boldsymbol{X_j^T y}|| < \lambda_n \text{ for } \beta_j = 0 \tag{5.3}$$

(5.2) and (5.3) are called the Karush-Kuhn-Tucker (KKT) conditions of Lasso. They are sufficient conditions. For necessary conditions, replace $<$ in (5.2) with $\leq$. There are indeed p conditions in total.

In general, the KKT conditions of claiming x^* is the optimizer of a function $f(x)$ is

$$\mathbf{0} \in \partial f(x^*)$$

5.c Local Quadratic Approximation

We will use the example of logistic regression. Suppose that

$$\log \frac{\Pr(y=1|x)}{\Pr(y=0|x)} = \beta_0 + x^T\beta$$

The log-likelihood of the logistic model is

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \big\{ y_i(\beta_0 + x_i^T\beta) - \log\big(1 + \exp\big(\beta_0 + x_i^T\beta\big)\big)\big\} \tag{5.4}$$

We want to solve

$$\hat{\beta}^{\text{lasso}} = \underset{\beta_0, \beta}{\arg\min} -\ell(\beta_0, \beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Suppose $(\tilde{\beta}_0, \tilde{\beta})$ is the current solution. We do a local quadratic approximation to the negative log-likelihood. Let $\ell_i(\beta_0, \beta)$ be the i-th summand on the right-hand side of (5.4) (i.e., ℓ_i is the log-likelihood based on only the i-th observation). We have

$$\begin{aligned} -\ell_i(\beta_0, \beta) \approx & -\ell_i(\tilde{\beta}_0, \tilde{\beta}) - \ell_i'(\tilde{\beta}_0, \tilde{\beta})\{\beta_0 + x_i^T\beta - \tilde{\beta}_0 - x_i^T\beta\} \\ & - \frac{1}{2}\ell_i''(\tilde{\beta}_0, \tilde{\beta})\{\beta_0 + x_i^T\beta - \tilde{\beta}_0 - x_i^T\tilde{\beta}\}^2 \end{aligned}$$

After some algebra, we find

$$\begin{aligned} -\ell_i'(\tilde{\beta}_0, \tilde{\beta}) &= y_i - \tilde{p}(x_i) \\ -\ell_i(\tilde{\beta}_0, \tilde{\beta}) &= -\tilde{p}(x_i)(1 - \tilde{p}(x_i)) \end{aligned}$$

where

$$\tilde{p}(x_i) = \frac{\exp\Big(\tilde{\beta}_0 + x_i^T\tilde{\beta}\Big)}{1 + \exp\Big(\tilde{\beta}_0 + x_i^T\tilde{\beta}\Big)}$$

Substituting these expressions in, we have

$$-\ell(\beta_0, \beta) \approx \frac{1}{2}\sum_{i=1}^{n} w_i(z_i - \beta_0 - x_i^T\beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})$$

where

$$\begin{aligned} z_i &= \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{p}(x_i))} \\ w_i &= \tilde{p}(x_i)(1 - \tilde{p}(x_i)) \end{aligned}$$

Therefore, at the current step, we only need to solve the following weighted least squares problem, which can be solved by the coordinate descent algorithm:

$$(\tilde{\beta}_0^{\text{new}}, \tilde{\beta}^{\text{new}}) = \underset{(\beta_0,\beta)}{\arg\min} \frac{1}{2}\sum_{i=1}^{n} w_i(z_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

The whole algorithm for solving this optimization problem with a fixed λ is as follows:

- Outer loop: Update the quadratic approximation ℓ_Q using the current parameters $(\tilde{\beta}_0, \tilde{\beta})$.
- Innter loop: Run the coordinate descent algorithm on the penalized weighted least squares problem.
- Iterate between the two loops.

The algorithm minics the Iteratively Reweighted Least Squares (IRLS) for solving the ordinary logistic regression (without penalty). IRLS also uses local quadratic approximation to the log-likelihood function at each iteration. The extra effort brought by penalization is essentially the soft-thresholding.

5.d Marjorization Minimization (MM) Algorithm

In the Coordinate Descent Algorithm for GLMs, the weights w_i depend on the current solution $(\beta_0, \tilde{\beta})$. To make computation easier, we can replace the weights with a scalar. In the logistic regression case, $w_i \leq 1/4$ (since the maximum $p(1-p)$ happens at $p = 1/2$. Thus, we can replace each weight w_i with $1/4$. This leads to the Majorization Minimization (MM) Algorithm.

The key idea of the MM algorithm is to find a function dominating the objective function from above, and then minimize the dominating objective function. Usually the minimizer of the dominating function needs to be easy to be solved.

Suppose $\ell(\beta)$ is twice continuously differentiable. We have

$$\begin{aligned}\ell(\beta) &= \ell(\beta) + [\nabla\ell(\tilde{\beta})]^T(\beta - \tilde{\beta}) + (\beta - \tilde{\beta})^T[\nabla^2\ell(\tilde{\beta})](\beta - \tilde{\beta}) \\ &\leq \ell(\tilde{\beta}) + [\nabla\ell(\tilde{\beta})]^T(\beta - \tilde{\beta}) + c(\beta - \tilde{\beta})^T(\beta - \tilde{\beta})\end{aligned}$$

where $c \geq \sup_\beta \lambda_{\max}(\nabla^2\ell(\beta))$. Then we can update $\tilde{\beta}$ by

$$\tilde{\beta}^{\text{new}} = \tilde{\beta} - \left\{\frac{1}{2c}[\nabla\ell(\tilde{\beta})]\right\}$$

5.e Properties of Desirable Penalty Functions

We want to design penalty function which should result in an estimator with the following three properties.

- **Unbiasedness:** the resulting estimator should be unbiased when the true unknown parameter is large to avoid unnecessary modeling bias. To check this, we need to show

$$p'_\lambda(|\theta|) = 0 \text{ for large } \theta$$

- **Sparsity:** the resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity. To check this, we show

$$p_\lambda(\cdot) \text{ should be singular at the origin.}$$

- **Continuity:** the resulting estimator is continuous in data z to avoid instability in model prediction. To check this, we show

$$|\theta| + p'_\lambda(|\theta|) \text{ attains its minimum at the origin}$$