The Concept of Stringency for Test Comparison: The Case of a Cauchy Location Parameter

Arif Zaman® — Lahore University of Management Sciences, Lahore

Asad Zaman — Pakistan Institute of Development Economics, Islamabad

Atiq ur Rehman — Pakistan Institute of Development Economics, Islamabad

Received: 09.03.2017 Accepted: 08.06.2017 Published: 03.04.2017
doi:10.33818/ier.319909

ABSTRACT

Strategies for comparison of alternative tests do not receive much attention in econometrics. The purpose of this paper is to introduce the concept of stringency and illustrate it in the context of a very simple hypothesis testing problem. Systematic use of this concept can be very helpful in evaluating relative performance of tests.

Key words: *Power Envelope, Location Parameter, Stringent Test*
JEL Classifications: C12

1. INTRODUCTION

We start by posing a puzzle: why did the Durbin and Watson (1950) and Durbin and Watson (1951) papers on testing for serial correlation become among the most highly cited econometrics papers of the era? It was clear to contemporaries that serial correlation was widespread in time series, but all tests necessarily depend on the structure of the matrix of regressors. The contribution of Durbin and Watson was to create bounds which were invariants across all design matrices, and hence *permitted tabulation of critical values*. Without such a table, limitations in computational capabilities made it impossible to test for serial correlation. Advances in computational capabilities have now made it a trivial matter to obtain simulated critical values for any test statistic, taking the design matrix into account. Contemporary journals would reject the Durbin Watson paper; it provides an unnecessary approximation.

While the concept of "stringency" is critical for the evaluation of tests, it receives virtually no attention in textbooks; see however, Zaman (1996) for an extended discussion. The main reason for this neglect is that the heavy computations required for widespread application of this concept have become possible only recently, due to massive increases in computational capabilities of computers.

Lehmann and Romano (2005, p. 245) write that methods based on invariance and unbiasedness apply to a restricted class of problems. Outside this class, the problem of what to look for in a good test is left not only without a solution, "but even without a formulation." Later, they

® Arif Zaman, Professor of Statistics and Mathematics, Lahore University of Management Science, Lahore, (email: arifz@lums.edu.pk)
Asad Zaman, Vice Chancellor, Pakistan Institute of Development Economics, Islamabad, (email: asadzaman@alum.mit.edu), Tel: +92-51-9248042, Fax: +92-51-9248025.
Atiq ur Rehman, Assistant Professor, Department of Econometrics and Statistics, Pakistan Institute of Development Economics, Islamabad, (email: atiq@pide.org.pk) Phone +92-51-9248060, Fax: +92-51-9248025

suggest that "stringency" provides a possible formulation, but it is difficult to compute the most stringent test when considerations of invariance cannot be applied. The goal of our paper is to show that advances in computational power have made it much more feasible to look for the most stringent test, although finding it remains computationally burdensome. Generations of statisticians brought up on this canonical text have absorbed the lesson that except in a very narrow class of problems, finite sample comparisons of hypothesis tests can only be done in an intuitive and ad-hoc way. Asymptotic theory does create substantial simplifications, so that globally valid comparisons of tests can be carried out, and several methodologies for doing this have been developed by a number of authors. However, asymptotic comparisons of tests suffer from the following as yet unresolved problems:

1. Three different ways to taking limits to calculate the asymptotic power lead to three different criteria for optimality of tests: Pitman efficiency, Bahadur efficiency, and Hodges-Lehmann efficiency. Unfortunately, these are often in conflict, and no clear resolution of these conflicts is available.

2. One way to resolve the conflict is to study how well the finite sample situation is approximated by the three different types of asymptotics. While there are many results in this direction, no clear conclusion has emerged from this line of research.

3. Asymptotics provide a coarse gauge of test performance. Large numbers of tests with vastly different finite sample properties turn out to be asymptotically equivalent for one or more of the asymptotic performance criteria.

In particular, asymptotic methods of comparison prove *LM* (Lagrange multiplier), Wald and *LR* (likelihood ratio) tests to be first order equivalent, but these can have drastically different finite sample performances. The results of this lack of a methodology for finite sample comparison of tests, is chaos. Analytics are generally impossibly complex, so simulation studies are the only feasible means for test comparisons. However, simulations studies show what is obvious a priori: each test has areas of strength and weakness. This means that tests cannot be compared without knowing the alternative hypothesis, but tests are useful only when we do not know whether or not the null hypothesis holds, and do not know the alternative. We can cite hundreds of studies with different and conflicting recommendations for hypothesis tests, since there is no standard method for comparisons; see Islam (2017) for illustrations of these conflicts.

Over the past few decades, massive increases in computational power have made feasible a computational approach to stringency which was not originally possible. This makes it possible to apply the concept to a much wider class of problems than those treated by Lehmann. We propose the use of "stringency" as a Gold Standard for the evaluation of tests. For a one-dimensional parameter, it should usually be possible to evaluate this one number numerically for most hypothesis testing problems. Zaman (1996) utilized this methodology to show that the popular Durbin-Watson test for autoregressive errors in regression model was very poor compared to certain alternatives. More recently, Khan (2017) has used the approach to compare tests of normality and come up with definitive recommendations. For higher dimensional problems, there remain formidable obstacles to the numerical evaluation of stringencies. Nonetheless, the concept sets up a clear target for what to look for in tests; this contrasts with Lehmann's pessimistic conclusion that the problem is left not only without a solution, but "even without a formulation." **Stringency provides a formulation of the problem of what to look for in a good test.** Even if exact evaluation of stringencies is not possible, a large number of strategies can be used to provide an approximation to this one number which provides a clear

cut evaluation and ranking of all tests. Having a goal, a well-defined target number we are trying to calculate as a single performance measure for all tests, would lead to substantial clarity even in situations where only approximations to it are available.

Even though we find occasional applications of it in the literature, the basic concepts involved in using stringency as a measure of finite sample performance of tests, remain unfamiliar to most. The goal of this paper is to provide an exposition of stringency in the context of a very simple example: tests for location parameter of a single draw from a Cauchy distribution. This example is chosen since many of the required tools can be analytically calculated, while those which cannot be, are easily evaluated numerically.

A quick review of the notation and definition of stringency and related concepts is given in Section 2. From Section 3, we specialize to the case of testing for a Cauchy location parameter. Section 4 describes the Neyman Pearson (*NP*) tests in this context. The power function of these tests is computed in Section 5. A qualitative description and graphs of the power envelope, the shortcoming, and the stringency for the Neyman Person tests are shown in sections 6 and 7 and 8 respectively. Section 8 then concludes by finding the most stringent *NP* test, followed by a discussion of why even more stringent tests would be possible if we did not limit our search to only *NP* tests.

2. CONCEPT OF STRINGENCY IN HYPOTHESIS TESTING

We review the basic principles of hypothesis testing in order to set up the notation, terminology and framework for our discussion of stringency. Suppose that we have a vector of observations X, which comes from a parametric family of densities $X \sim f(x,\theta)$. The parameter $\theta \in \Theta$ is an element of the parameter space Θ.

2.1. Hypotheses

A hypothesis test consists of two mutually disjoint subsets Θ_0 and Θ_1 of Θ that are interpreted as

$$\begin{aligned} \text{Null Hypothesis} \quad & \mathcal{H}_0 \colon X \sim f(x,\theta) \text{ for some } \theta \in \Theta_0 \\ \text{Alternative Hypothesis} \quad & \mathcal{H}_1 \colon X \sim f(x,\theta) \text{ for some } \theta \in \Theta_1 \end{aligned}$$

2.2. Tests and Rejection Regions Hypotheses

Any function $T(X)$ taking values $\{0,1\}$ is called a test[1], with the interpretation that when X is observed, we accept the null hypothesis if $T(X) = 0$ and reject if $T(X) = 1$. The set of all values of X for which $T(X) = 1$ is called the rejection region of the test, and tests can alternatively be characterized by their rejection regions.

2.3. Power, Size and Errors of Types I and II

There are two types of errors that can occur during hypothesis testing, and the probabilities of these errors can be used to evaluate the performance of tests. A type II error is when the Null Hypothesis is actually false, but the test does not reject the Null Hypothesis. The power of a

[1] Randomized tests can take intermediate values, but these can be ignored for the purposes of the present discussion. Our goal is to present an exposition in the simplest possible framework.

test is one minus the probability of type II error, which is a function of the exact value of θ in the parameter space of alternatives Θ_1. Using Π to denote power,

$$\text{Power of } T(\cdot) = \Pi(T,\theta) = 1 - P(T(X) = 0 \mid \theta) = P(T(X) = 1 \mid \theta) \tag{2.1}$$

where the domain of θ is considered limited to the Θ_1 region. A type I error is when the Null Hypothesis is true, but the test rejects the Null Hypothesis. The size (also called level) of a test is defined as the maximum possible probability of type I error:

$$\text{Size of } T(\cdot) = \mathcal{L}(T) = \sup_{\theta} P(T(X) = 1 \mid \theta \in \Theta_0) \tag{2.2}$$

$$= \sup_{\theta \in \Theta_0} \Pi(T,\theta) \tag{2.3}$$

Note the abuse of notation in the last equation, where we are using the power function Π outside of its usual domain. For a test to be considered good it should have a small size and a high power.

2.4. Power Envelope

Let us fix the size, α, of the test and define $\mathcal{T}_\alpha$ be class of all tests of size α. To define stringency, the crucial concept is the power envelope, which is the maximum possible power that can be achieved at a given alternative. For any given size α, this can be defined as follows:

$$MP(\theta,\alpha) = \sup_{T \in \mathcal{T}_\alpha} \Pi(T,\theta) \tag{2.4}$$

This is the maximum possible power achievable against the alternative θ by any test T of size α.

2.5. Shortcoming

The shortcoming, S, of any test T is measured by its performance relative to the power envelope:

$$S(T,\theta) = MP(\theta,\alpha) - \Pi(T,\theta) \tag{2.5}$$

A test with zero shortcoming at θ is called the Most Powerful test for the alternative θ. Such a test has the property that no other test of equal size can have more power at θ.

2.6. Stringency

The stringency of a test is the maximum shortcoming of a test evaluated over the entire space of alternatives:

$$S(T) = \max_{\theta \in \Theta_1} S(T,\theta) \tag{2.6}$$

This is a single number which measures the overall performance of the test. This means that all tests can be compared and ranked on the basis of this measure. Furthermore, the stringency of a test has a natural and intuitive explanation. A test with stringency zero is a uniformly most powerful test – it is most powerful for all alternatives. This test should always be preferred, if it exists. If a test has stringency of 1%, it means that the test has power only 1% less than the most powerful test available at all possible alternatives. For practical purposes, this test is nearly as good as a uniformly most powerful test. If tests with stringencies between 5% to 10% can be found, then we need search no further for practical purposes. If the best available tests have stringency of 50% or more, then we should search for better methods of testing. The point is that evaluating stringency of tests provides us with a very important guide to the use and

comparison of tests in practical problems. Previously, this evaluation was only possible in a very narrow class of problems, and hence the concept of stringency was ignored as being of limited practical value. Massive increases in computer power have made possible to evaluation of stringency in a much larger class of problems. The goal of this article is to illustrate this possibility.

3. NEYMAN-PEARSON TESTS FOR CAUCHY LOCATION PARAMETER

To clarify the concepts discussed so far, we illustrate them all within the context of an example. Suppose X is a random variable with a Cauchy distribution, with location parameter θ: $X \sim (\theta)$. The likelihood function for X, which is also the density of X, is:

$$l(x,\theta) = \frac{1}{\pi[1+(x-\theta)^2]}$$

Throughout this paper, we consider the one-sided problem of finding a hypothesis test in the following situation

$$\begin{aligned} \text{Null Hypothesis} \quad & \mathcal{H}_0\text{: } X \sim \mathcal{C}(0). \\ \text{Alternative Hypothesis} \quad & \mathcal{H}_1\text{: } X \sim \mathcal{C}(\theta) \text{ for some } \theta > 0 \end{aligned} \tag{3.7}$$

We occasionally want to consider only a single point alternative, in which case we will use the notation

$$\text{Point Alternative} \quad \mathcal{H}_\theta\text{: } X \sim \mathcal{C}(\theta) \text{ for some } \theta > 0$$

According to the Neyman Pearson lemma, the most powerful test against a point alternative $\theta \in \Theta_1$ must reject the Null hypothesis for all x for which the likelihood ratio

$$LR(x,\theta) = \frac{l(x,0)}{l(x,\theta)} \le C \tag{3.8}$$

for some constant C. While the usual definition of the likelihood ratio is the inverse of our definition in equation 3.8, we will use the reversed ratio, because it simplifies the algebra. For the Cauchy problem at hand, this can be written explicitly as

$$LR(x,\theta) = \frac{1+(x-\theta)^2}{1+x^2} \tag{3.9}$$

With this definition, all Neyman Pearson tests have rejection regions of the form

$$RR(\theta,C) = \{x: LR(x,\theta) \le C\} = \{x: \frac{1+(x-\theta)^2}{1+x^2} \le C\} \tag{3.10}$$

$$= \{x: 0 \le (C-1)x^2 + 2\theta x + (C-1-\theta^2)\} \tag{3.11}$$

Neyman Pearson tests are all the indicator functions of rejection regions

$$NP(X,\theta,C) = I(X \in RR(\theta,C))$$

The size or level of any such test is

$$\begin{aligned} \alpha &= \mathrm{E}(NP(X,\theta,C) \mid \mathcal{H}_0) \\ &= P(X \in RR(\theta,C) \mid \mathcal{H}_0) \\ &= P(LR(X,\theta) \le C \mid \mathcal{H}_0) \end{aligned}$$

The most powerful test against a point alternative θ_1 is $NP(x,\theta_1,C^*(\theta_1,\alpha))$, with fixed level α, where $C^*(\theta_1,\alpha)$ is the value of C for which the level is α. In other words,

$$P(LR(X,\theta_1) < C^*(\theta_1,\alpha) \mid \mathcal{H}_{\theta_1}) = \alpha \tag{3.12}$$

We often wish to consider the α-level Neyman-Pearson test against a point alternative θ_1, and evaluate its performance (power) when the actual value of the alternative is θ. We will denote this function of three variables as

$$\begin{aligned} \Pi_{NP}(\theta_1,\alpha,\theta) &= \Pi(NP(X,\theta_1,C^*(\theta_1,\alpha)),\theta) \\ &= \mathrm{E}(NP(X,\theta_1,\ C^*(\theta_1,\alpha)) \mid \mathcal{H}_\theta) \\ &= P(X \in RR(\theta_1,\ C^*(\theta_1,\alpha)) \mid \mathcal{H}_\theta) \\ &= P(LR(X,\theta_1) \le C^*(\theta_1,\alpha) \mid \mathcal{H}_\theta) \end{aligned} \tag{3.13}$$

Finally, because Neyman Pearson tests, $NP(X,\theta,C)$, are the most powerful tests against the point alternative θ, we can claim that

$$MP(\theta,\alpha) = \Pi_{NP}(\theta_1,\alpha,\theta) \tag{3.14}$$

is the maximum power possible for any α-level test, where C^* is as defined in equation 3.12.

3.1. Explanation of the Neyman Pearson Lemma

The first goal in the calculation of stringency is the calculation of the power envelope. This is a function of θ_1 that maps each value in the alternative to the maximum possible power that attainable at θ_1 within the class of level α tests. In this particular problem, and in all one parameter problems, the test $NP(X,\theta,C)$ is the best possible test at the alternative θ among all tests having the same significance level as the *NP* test. The meaning of "best" is defined by the following theorem.

Theorem 3.1. (Neyman-Pearson) Suppose that the test $NP(X,\theta,C)$ has level α and $T(X)$ is a different test with level less than or equal to α. Then the test T must have less power than *NP* at the alternative θ.

In other words

$$\Pi(T,\theta) = P(T(X) = 1 \mid \mathcal{H}_\theta)) \le P(NP(X,\theta,C) = 1 \mid \mathcal{H}_\theta) = \Pi(NP(X,\theta,C),\theta)$$

Explanation: The mathematical proof is available from many sources. Here we offer an intuitive explanation. While the mathematical statement appears complex, the intuition is quite straightforward. The *NP* test is the best possible test because it rejects the null at all the points in the rejection region $RR(\theta,C) = \{X: LR(X,\theta) \le C\}$. Any different test must remove some points from this region, and replace them by points outside this region to get equal size. But points outside this region have $LR(X,\theta) > \mathrm{C}$, so that the null hypothesis is *more* likely at those points, while the alternative is *less* likely. Rejecting the null hypothesis when it is more likely leads to a loss in power.

Before doing explicit analytical calculations, it is useful to do a graphical analysis of the likelihood ratio, so as to acquire some intuitive understanding of the shape of the rejection regions for varying values of C and θ. This is undertaken in the next section.

4. REJECTION REGIONS OF THE MOST POWERFUL TEST

4.1. The Likelihood Ratio

Fixing the value of θ at 0.5, a graph of $LR(x,\theta = 0.5)$ is shown in Figure 4.1. The shaded area in the graph represents the rejection region.

For any $\theta > 0$ it can be verified by calculus that $LR(x,\theta)$ as defined in Equation 3.9 has the following properties:

1. $\lim_{x\to+\infty} LR(x,\theta) = 1 = \lim_{x\to-\infty} LR(x,\theta)$

2. $LR(0,\theta) = 1+\theta^2 > (LR(\theta/2,\theta) = 1) > LR(\theta,\theta) = \dfrac{1}{1+\theta^2}$

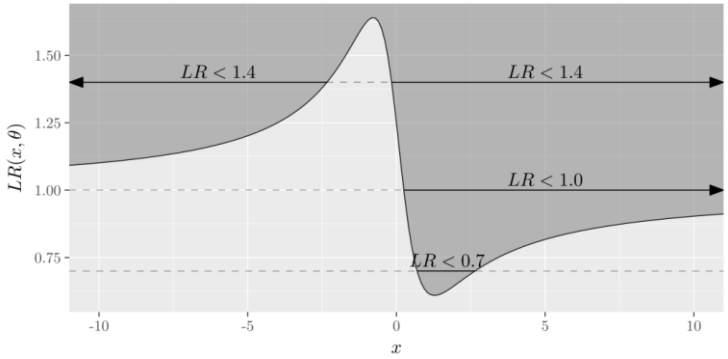
3. LR is continuous, having a minimum at $(x_{\min}(\theta), y_{\min}(\theta))$ and a maximum at $(x_{\max}(\theta), y_{\max}(\theta))$ with $x_{\min}(\theta) < 0 < x_{\max}(\theta)$,

4. The curve is strictly increasing from $-\infty$ to $x_{\max}$

5. The curve is strictly decreasing from $x_{\max}$ to $x_{\min}$

6. The curve is strictly increasing from $x_{\min}$ to ∞.

Figure 4.1 The $LR(x,\theta = 0.5)$ and Rejection Regions



Any horizontal line with height C between y_{min} and y_{max} (we use the shorthand x_{min} to refer to $x_{min}(\theta)$, when there is no danger of confusion) will hit this curve once at some point $x_{\text{in}}(\theta,C)$ between x_{min} and x_{max}. If $C \neq 1$, there will be exactly one other point of intersection at $x_{out}(\theta,C)$. This second point will be to the right of x_{max} if $C < 1$ and to the left of x_{min} if $C > 1$.

4.2. Rejection Regions in terms of Likelihood Ratio

Given any value of C between y_{min} and y_{max}, we can characterize three different kinds of rejection regions.

1. If $C < 1$ then $x_{\text{in}} < x_{\text{out}}$ and $RR(\theta,C) = (x_{\text{in}}, x_{\text{out}})$.

2. If $C = 1$ then x_{out} does not exist and $RR(\theta,C) = (x_{\text{in}}, \infty)$.

3. If $C > 1$ then $x_{\text{out}} < x_{\text{in}}$ and $RR(\theta,C) = (x_{\text{in}}, x_{\text{out}})^c$ which is the complement of an interval.

The actual coordinates of the extrema are given by

$$(x_{\min}, y_{\min}) = \left(\frac{\theta + \sqrt{\theta^2+4}}{2}, x_{\max}^2 \right)$$

$$(x_{\max}, y_{\max}) = \left(\frac{\theta - \sqrt{\theta^2 + 4}}{2}, x_{\min}^2 \right)$$

For values of $C < y_{min}$ the rejection region is the empty set, and for $C > y_{max}$, the entire real line is the rejection region. Between these two limits, the rejection region grows monotonically as C increases. Correspondingly, the probability $P(RR(\theta, C) | \mathcal{H}_\theta)$ increases continuously from 0 to 1 as C increases. Given any level $0 < \alpha < 1$, there is a unique value $C^*(\alpha, \theta)$ at which the test $NP(X, \theta, C^*(\alpha, \theta))$ has level α, and that value is obtained by solving the equation $P(RR(\theta, C) | \mathcal{H}_\theta) = \alpha$ for C. As this is a Neyman Pearson test, amongst all tests of level α it is the most powerful test.

4.3. The Level of a *NP* Test

The most powerful level α test is characterized by the following theorem:

Theorem 4.2. (α level *NP* Test Neyman-Pearson): The most powerful level α test for the Hypotheses in (3.7) rejects the null for all $x \in RR(\theta, C^*(\theta, \alpha))$ where

$$C^*(\theta, \alpha) = 1 + \frac{\theta^2}{2} - \frac{\theta \theta_\alpha}{2} \sqrt{\frac{\theta^2 + 4}{\theta_\alpha^2 + 4}} \quad \text{and} \quad \theta_\alpha = \frac{2}{\tan \pi\alpha} \tag{4.15}$$

The definition of θ_α is motivated by the fact that it is that value of θ for which $C^*(\theta, \alpha) = 1$. Another way of saying it is that the test $NP(x, \theta_\alpha, 1)$ has level α. Or using plainer words, when the level of the test is set to α, and the likelihood ratio cutoff is set to $C = 1$, then θ_α is the value of the alternative θ against which a test has level α.

Proof: For the case of $C = 1$, it is easy to see that the rejection region of the Neyman Pearson test $NP(x, \theta, 1)$ is of the form $RR(\theta, 1) = (\theta/2, \infty)$, and that its level is

$$\int_{\theta/2}^{+\infty} \frac{dx}{\pi(1 + x^2)} = \frac{\arctan x}{\pi} \Big|_{\theta/2}^{+\infty} = \frac{\pi/2 - \arctan(\theta/2)}{\pi}$$

When we set the level at α, this becomes $\alpha = 1/2 - \arctan(\theta/2)/\pi$. When we solve this for θ, the result is $\theta_\alpha = 2 / \tan(\pi\alpha)$, which proves the theorem when $C = 1$.

When $C \neq 1$, the two points where the horizontal line at C meets the LR curve are the roots of the equation $LR(x, \theta) = C$ or

$$(1 - C)x^2 - 2\theta x + 1 - C + \theta^2 \tag{4.16}$$

They can be computed to be

$$x_{in} = \frac{\theta + \sqrt{C\theta^2 - (1 - C)^2}}{1 - C} \tag{4.17}$$

$$x_{in} = \frac{\theta - \sqrt{C\theta^2 - (1-C)^2}}{1-C}$$

Note that when ($C < 1$) then $x_{\text{in}} < x_{\text{out}}$, and when $C < 1$ the inequality is reversed. Furthermore, when $C < 1$ the leading term of the quadratic equation 4.16 is positive, so the smaller values (which comprise the rejection region) will be between the roots, and when $C > 1$ the leading term is negative so the smaller values will occur outside the roots.

The level of the *NP* test, when $C < 1$ is given by

$$\int_{x_{in}}^{x_{out}} \frac{dx}{\pi(1+x^2)} = \frac{\arctan x}{\pi}\Big|_{x_{in}}^{x_{out}} \tag{4.18}$$

When $C > 1$ the value $x_{\text{out}} < x_{\text{in}}$ so this integral in equation 4.18 is negative. Also in this case, we are looking for the probability outside the interval, so the level of the *NP* test for this case is given by

$$\int_{x_{in}}^{x_{out}} \frac{dx}{\pi(1+x^2)} + 1 = \frac{\arctan x}{\pi}\Big|_{x_{in}}^{x_{out}} + 1 \tag{4.19}$$

Actually, the two cases in the previous equation are not really different. The arctan is a multi-valued function, of which we usually take the principal branch, which take values between $-\pi/2$ and $\pi/2$. If instead, we choose the discontinuous branch taking values between 0 and π, the "+1" would occur automatically. We will instead, rewrite the above two equations as one by simply adding an indicator function as

$$\int_{x_{in}}^{x_{out}} \frac{dx}{\pi(1+x^2)} + 1 = \frac{\arctan x}{\pi}\Big|_{x_{in}}^{x_{out}} + I\{x_{out} < 0\} \tag{4.20}$$

If we set the level equal to α we must have

$$\arctan x_{out} - \arctan x_{in} = \pi(\alpha - I\{x_{out} < 0\})$$

Since tan is has a period of π, if we take tan of both sides, $\tan(\pi\alpha) = \tan(\pi\alpha - \pi)$, so we can ignore the indicator function. On the left hand side, we can use the trigonometric identity $\tan(a-b) = (\tan(a) - \tan(b))/(1+\tan(a)\tan(b))$ to get

$$\frac{x_{out} - x_{in}}{1 + x_{out}x_{in}} = \tan \pi\alpha$$

Substituting the roots from equation 4.17 and using the definition of θ_α in the statement of this theorem yields

$$2/\theta_\alpha = \tan \pi\alpha = \frac{2\sqrt{C\theta^2 + (1-C)^2}}{\theta^2 + 2(1-C)} \tag{4.21}$$

Squaring both sides, and rearranging

$$\theta_\alpha^2 (C\theta^2 - (1-C)^2) = (2(1-C) + \theta^2)^2$$

This last is a quadratic equation in C with two roots, which, after some simplification turn out to be

$$C = 1 + \frac{\theta^2}{2} \pm \frac{\theta\theta_\alpha}{2}\sqrt{\frac{\theta^2+4}{\theta_\alpha^2+4}} \tag{4.22}$$

Since $y_{max} = 1+\theta^2/2+(\theta\sqrt{\theta^2+4})/2$, only the smaller root in the above equation is valid as a solution, which is the claim of the theorem.

This is the proof for all cases. □

For any fixed point alternative θ, equation 4.22 allows us to solve for the likelihood ratio cutoff of the most power test, $C = C^*(\theta,\alpha)$ in terms of α. Similarly, equation 4.21 can easily be solved to give the level α of the most powerful test in terms of the likelihood ratio cutoff C.

We can then compute the limits of the rejection region, x_{in} and x_{out} by plugging the value of $C^*(\theta,\alpha)$ and θ_α from the theorem into equation 4.17. After some simplification this results in

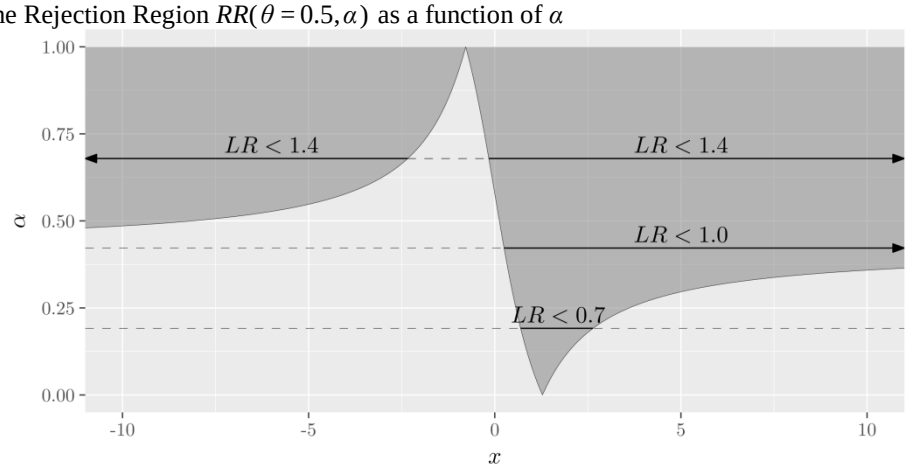
$$x_{in}(\theta,\alpha) = \frac{2+\sin(\pi\alpha)\sqrt{\theta^2+4}}{-\theta+\cos(\pi\alpha)\sqrt{\theta^2+4}} \tag{4.23}$$

$$x_{out}(\theta,\alpha) = \frac{2-\sin(\pi\alpha)\sqrt{\theta^2+4}}{-\theta+\cos(\pi\alpha)\sqrt{\theta^2+4}} \tag{4.24}$$

Theorem 4.3. Using the definitions given in equation 4.23, the rejection region for the most powerful α level test is given by

$$RR(\theta,\alpha) = \begin{cases} (\theta_\alpha/2,\infty) & \text{if} \quad \theta = \theta_\alpha \\ (x_{in}(\theta,\alpha), x_{out}(\theta,\alpha)) & \text{if} \quad 0 < x_{out}(\theta,\alpha) \\ (-\infty, x_{out}(\theta,\alpha)) \cup (x_{in}(\theta,\alpha),+\infty) & \text{if} \quad 0 > x_{out}(\theta,\alpha) \end{cases}$$

Figure 4.2 The Rejection Region $RR(\theta = 0.5,\alpha)$ as a function of α



4.4. Rejection Regions in terms of the α-Level

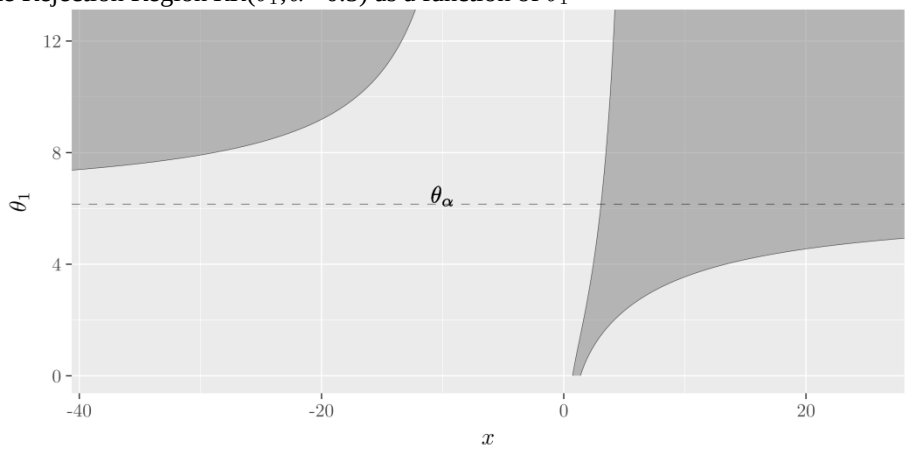
To get an idea of the structure of these tests, below are shown graphs of the rejection region with the value of θ fixed at 0.5.

Note that this is exactly the same as the graph in Figure 4.1, except that the *y*-axis has undergone a nonlinear monotonic transformation, given by the maps between $C^*(\theta,\alpha)$ and α given in equations 4.15 and 4.21.

Note also that for any particular observed value X, the *p*-value (or the observed significance level), is defined as the smallest α for which X lies in the rejection region. In fact, the curve in Figure 4.2 is a graph of *p*-value on the *y*-axis against observations X on the *x*-axis.

Another way to look at the rejection regions is to fix a level α at a particular value, say 0.05, and see how the rejection region varies as a function of the alternative θ_1. This is shown in Figure 4.3.

Figure 4.3 The Rejection Region $RR(\theta_1,\alpha=0.5)$ as a function of θ_1



5. THE POWER FUNCTION OF *NP* TESTS

Remember that the power of a test T is a function of θ as defined in equation 2.1

The left graph in Figure 5.4 is the power function

$$\Pi_{NP}(\theta_1=2,\alpha=2,\theta)$$

of the fixed test: $NP(X,\theta_1=2,\alpha=0.2)$. It is often a point of confusion, that the maximum of this power function does not occur at $\theta=2$. In fact, the power function often fails to have a maximum, because it often approaches 1 in the limit as $\theta\to\infty$. This is because a power function takes one fixed test, designed to be optimal at the point $\theta_1=2$ and measures its power at other θ_s for which the test was not designed.

As compared to the power function, if we were to fix the level α and alternative θ where the power is to be evaluated, and consider all possible tests, we would find that the test $NP(X,\theta_1=\theta,\alpha)$ would be the most powerful (this is exactly the result of the Neyman Pearson Lemma). In the right graph of Figure 5.4 the power of all Neyman Pearson tests against different alternatives θ_1 is shown when measured at $\theta=0.2$.

Both of the functions shown in Figure 5.4 are actually different views of the $\Pi_{NP}(\theta_1,\alpha,\theta)$ function defined in 3.13. The fact that it would take four dimensions to draw this function makes it challenging to visualize. Three of the four dimensions are depicted in Figure 5.5, which fixes α at the values of 0.05 and 0.20 and shows the other three dimensions. The curves of Figure 5.4 can be seen as two different slices from the image on the top right in Figure 5.5.

Figure 5.4 Power as a function of θ and θ_1

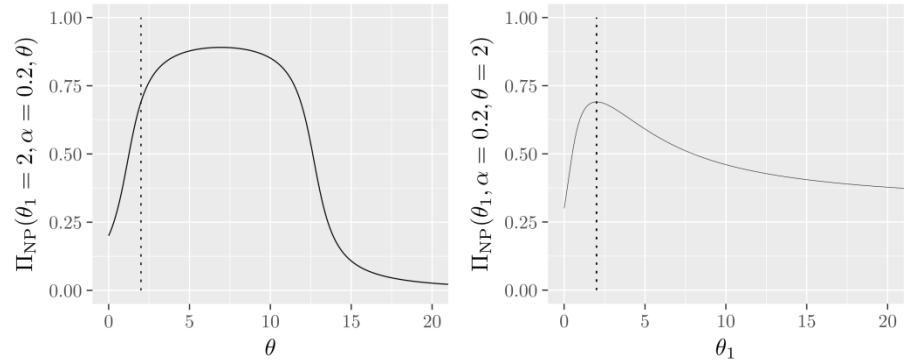
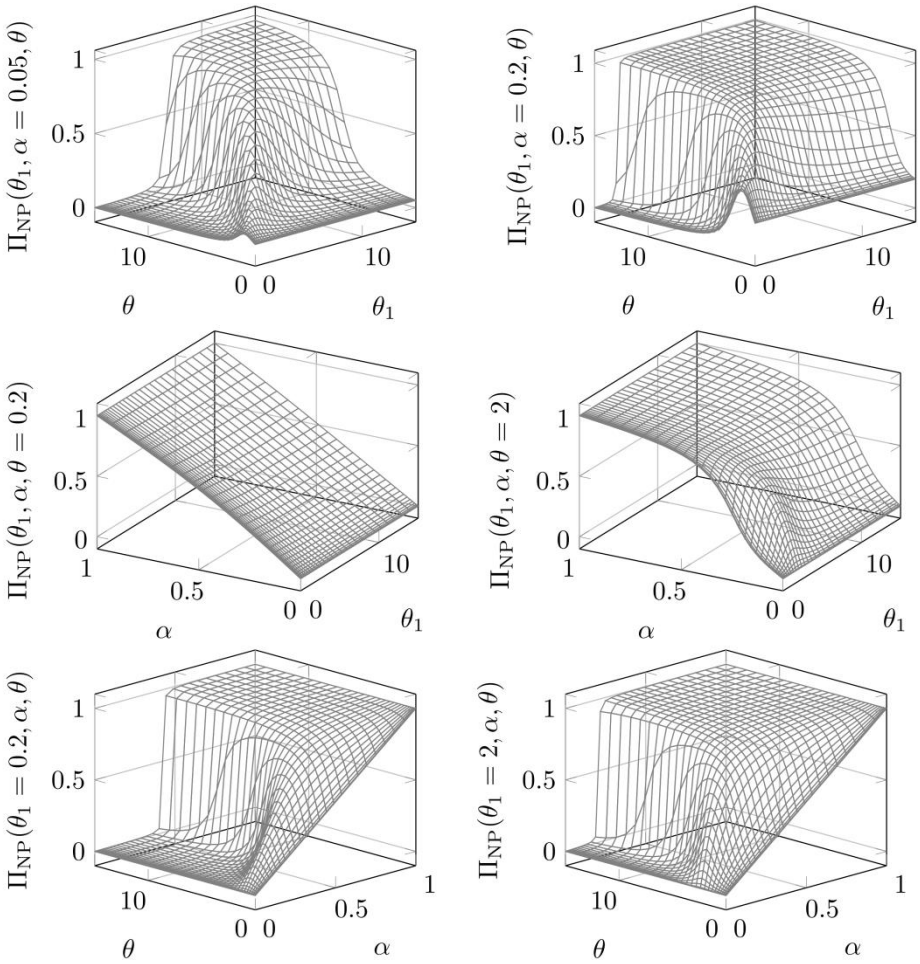


Figure 5.5 Power as a funtion of θ and θ_1



An obvious observation from the lower graphs of Figure 5.5 is that when $\theta = 0$ the power is equal to α. What is more interesting is the meaning of the power vs. α curve when $\theta_1 = 0$. When $\theta_1 = 0$, the likelihood ratio is simply a constant 1, so there are technically only two *NP* tests, One with $\alpha = 0$ which rejects nothing and one with $\alpha = 1$ which rejects everything. The shown curve is the limit of *NP* tests, as $\theta \to 0^+$. In this case, the rejection region is an interval, which can be seen graphically in Figure 5.3, and according to equation 4.23 works out to be $\sec(\pi\alpha) \pm \tan(\pi\alpha)$, and the probability of this region when the actual distribution is centered at θ is shown in the graph (when $\alpha > 0.5$ the tan becomes negative and we consider the outside of the interval rather than the inside, as previously).

6. THE POWER ENVELOPE

Knowing the power of the most powerful tests, it is easy to calculate the power envelope, which is the maximum possible power attainable at any alternative θ. Note that even though the power envelope was computed using only *NP* tests, it is the maximum possible power against all possible tests.

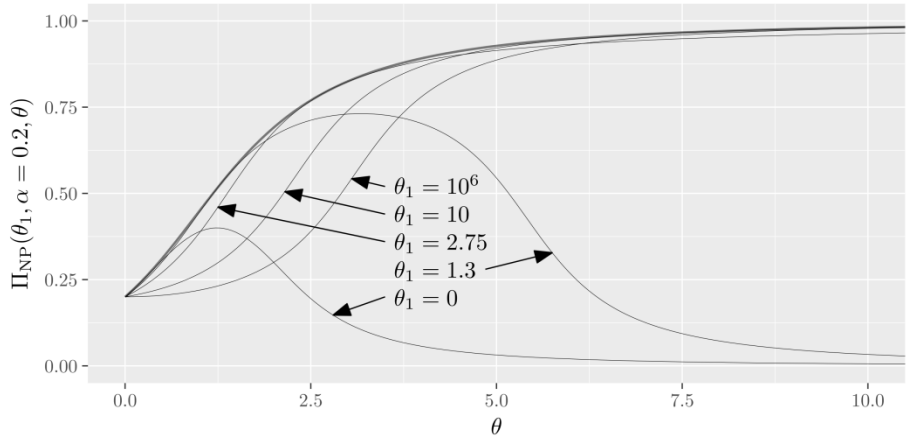
As mentioned in 2.4, the curve along the diagonal, $\Pi_{NP}(\theta,\alpha,\theta)$, is exactly the maximum power function, *MP*. It can be computed explicitly as

$$\begin{aligned} MP(\theta,\alpha) &= \Pi_{NP}(\theta,\alpha,\theta) \\ &= P(RR(\theta,\alpha)\,|\,\theta) = \int_{x_{in}(\theta,\alpha)}^{x_{out}(\theta,\alpha)} \frac{dx}{\pi(1+(x-\theta)^2)} \\ &= (\arctan(x_{out}(\theta,\alpha)-\theta) - \arctan(x_{in}(\theta,\alpha)-\theta))/\pi \end{aligned}$$

where the values of arctan are to be taken so that the answer ends up between 0 and 1, by reasoning similar to the explanation of equations 4.18 and 4.19. We already have formulas for x_{out} and x_{in} in equation 4.23.

The power function (the left curve of Figure 5.4) has been redrawn for many different values of θ_1 in Figure 6.6. The curves represent the values $\theta_1 = 0, 1.3, 2.75, 10$ and 10^6. A bold curve has been drawn to show the envelope of all these curves and this is the MP function. Note that each power curve tangentially touches the MP curve at the point $\theta = \theta_1$, a fact that is explained by equation 3.14.

Figure 6.6 Power Envelope and Power Functions for *NP* tests



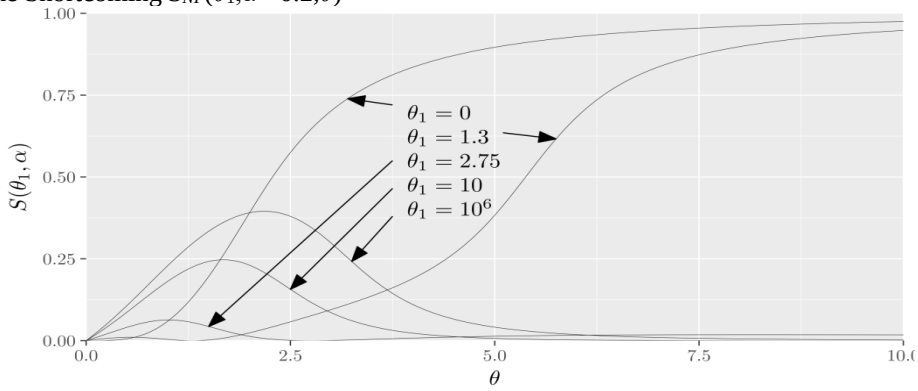
7. SHORTCOMING

The difference between the power envelope and the power function of any test is the shortcoming of that test. As we are only considering *NP* tests at the moment, we can introduce the notation

$$S_{NP}(\theta_1,\alpha,\theta) = S(NP(X,\theta_1,C^*(\alpha)),\theta)$$

to refer to the shortcoming of the α-level *NP* test against the alternative θ_1. If we subtract each of the power functions of Figure 6.6 from the power envelope, we can see the shortcoming of the tests more clearly in Figure 7.7.

Figure 7.7 The Shortcoming $S_{NP}(\theta_1,\alpha=0.2,\theta)$



There two different kinds of power curves in Figure 7.7 the ones with small values of θ_1 approach 1 in the limit as $\theta \to \infty$. The reason for this can be seen by noticing that if $\theta_1 < \theta_\alpha$, then the rejection region is bounded. As θ gets larger, the rejection region becomes infinitely far from θ, so the rejection probability falls to zero (which implies that the shortcoming approaches 1). On the other hand, as soon as $\theta_1 > \theta_\alpha$, the rejection region contains all values above x_{out}, so as θ gets large, the probability of the rejection region approaches 1 so the shortcoming approaches zero. Since $\theta_\alpha < 0$ for all $\alpha > 0.5$, so for large α, all power curves will be off the second type with the shortcoming approaching 0 as θ gets large.

Note that in Figure 6.6, $\theta_{\alpha=0.2} \approx 2.75$, which is the first curve that appears to approach 1 in the limit (actually, since $\theta_{\alpha=0.2}$ is slightly larger than 2.75, the drawn curve would eventually fall).

An α-level Neyman-Pearson test against the alternative θ_1 has zero shortcoming

- at $\theta = 0$, because at that point the level α is equal to the power.
- at $\theta = \theta_1$ because at that point the test is most powerful, and hence tangent to the power envelope.
- and in the limit as $\theta \to \infty$, when $\theta_\alpha < \theta_1$, as explained above.

7.1. A Conjecture

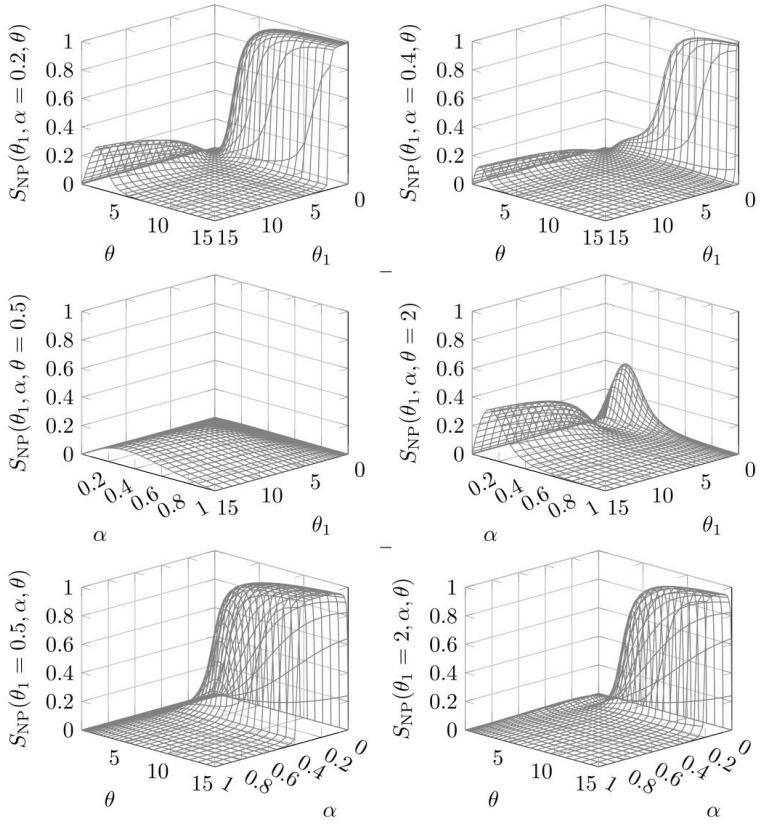
For the values of $\theta \geq \theta_\alpha$, from the graph it seems like the largest difference between MP and the power function occurs at small values of θ. One might conjecture that the test against $\theta = \theta_\alpha$ might be the test that achieves minimum shortcoming. We will see that while this conjecture is often true, it does not hold for all values of α.

7.2. The Shortcoming Function

As the shortcoming is also a function of three variables, $S(\theta_1,\alpha,\theta)$, we can look at three dimensional slices of this function by fixing one of the variables at some constant values.

The top left graph in Figure 7.8 shows a more complete picture of the same shortcoming shown in Figure 7.7. For any fixed value of θ_1, you can see the shortcoming curves of Figure 7.7.

Figure 7.8 Views of $S_{NP}(\theta_1, \alpha, \theta)$ with one variable fixed. The *z*-axis label identifies the variable held constant.



8. STRINGENCY

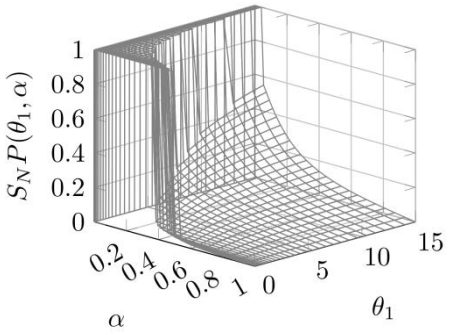
For any value of θ_1 and α, the stringency $S(\theta_1, \alpha)$ was defined to be the maximum shortcoming over all values of θ. While finding a global maximum can be a difficult problem in general, in our particular case, the simple structure of the shortcoming curves makes it quite easy, because there are two possible regions where the maximum might be, and each region has a simple maximum without any other apparent local maxima.

We are looking for the maximum shortcoming. While a general search for a maximum can often be difficult on a computer. Our problem is much simpler, because of the location of the zeros of the shortcoming function. There are two distinct regions, the first being between zero and θ_α, and the second region being all the values greater than θ_α. In the first region, the shortcoming seems to rise from zero to a maximum value and then fall back to zero. In the second region once again it rises, and then if $\theta_1 < \theta_\alpha$, it falls back approaching zero in the limit.

Since stringency is a function of two variables, so it can be summarized by just a single three dimensional graph of Figure 7.8. This graph has been obtained by doing a numerical maximization separately in the two regions mentioned in the previous paragraph.

A number of features that were anticipated can be clearly seen here. The stringency is 1, whenever $\theta_1 < \theta_\alpha$. Outside this region, for any fixed α, the stringency graph seems to be increasing in θ, so the conjecture made in section 7 appears to be holding.

Figure 8.9 A graph of the Stringency function $S(\theta_1, \alpha)$.



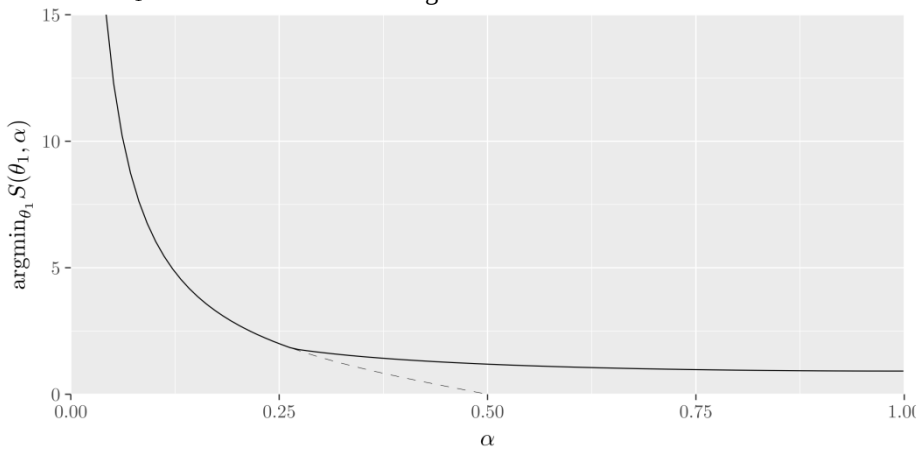
8.1. Most Stringent Tests

Stringency is our single standard by which we are measuring the quality of a test. For any given α, we want to find the most stringent *NP* test, i.e. the one that has the minimum stringency. From our conjecture, the minimum seems to be at $\theta_1 = 0$ for $\alpha \geq 0.5$ and for $\theta = \theta_\alpha$ for $\alpha < 0.5$. Figure 8.9 shows the value of θ_1 that achieves the minimum value (obtained numerically). The dotted line shows the conjectured value θ_α. As can be seen, the conjecture was good for $\alpha < 0.27$ (in fact the cutoff seems to be around 0.2671), for larger values of α, the minimum occurs somewhere else. We did not notice this by looking at the graphs because the values of stringency in that area are all very near zero, so difficult to distinguish by simply looking. We will look a bit more closely to see exactly what happens near $\alpha = 0.27$ that causes the sudden change in the next section.

The following graph shows the stringency of the most stringent Neyman Pearson Test in Figure 8.10. You can notice the kink near $\alpha = 0.27$ reflected in this graph as well.

It is important to note that this is *not* the most stringent test possible. As we mention in the next section, there are other non-Neyman-Pearson tests that can be even more stringent.

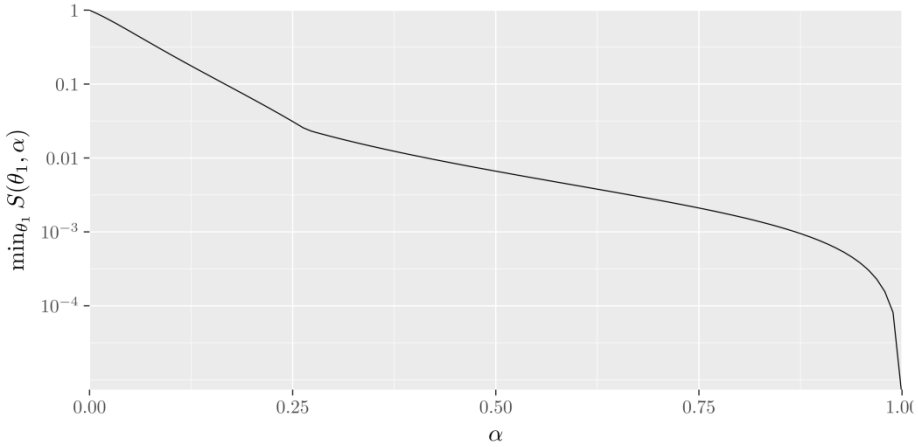
Figure 8.10 Location of θ_1 that attains the Most Stringent *NP* test.



8.2. The Change Near $\alpha = 0.27$

The central thick curve in Figure 8.12 is the graph of the shortcoming for $\alpha = 0.267$, and $\theta_1 = 1.8 = \theta_{\alpha=0.267}$. As mentioned before, shortcoming curves have two possible locations for the maximum. For this particular curve, both maxima are equal. The dashed line is for values of α 0.01 smaller and the dotted line for α being 0.01 larger than 0.267 (and θ_1 taking the corresponding θ_α values.

Figure 8.11 Stringency of the Most Stringent *NP* α-level Test.



As shown in Figure 8.9, for values of $\alpha < 0.267$, the θ_1 that minimizes the maximum stringency is θ_α , so the dashed line has the smallest shortcoming for that α. On the other hand, when $\alpha > 0.267$, increasing the value of θ_1 to a value larger than θ_α will reduce the second peak near $\theta = 6$ and increase the peak near $\theta = 0.5$, so that the overall maximum becomes smaller.

Figure 8.12 Shortcoming…

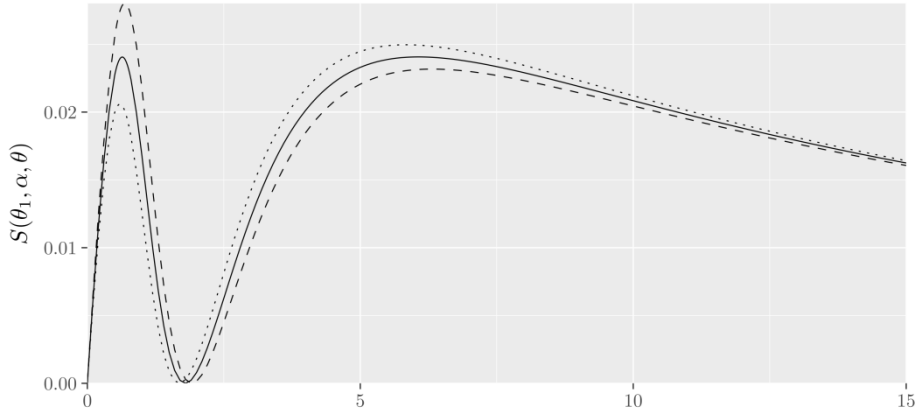
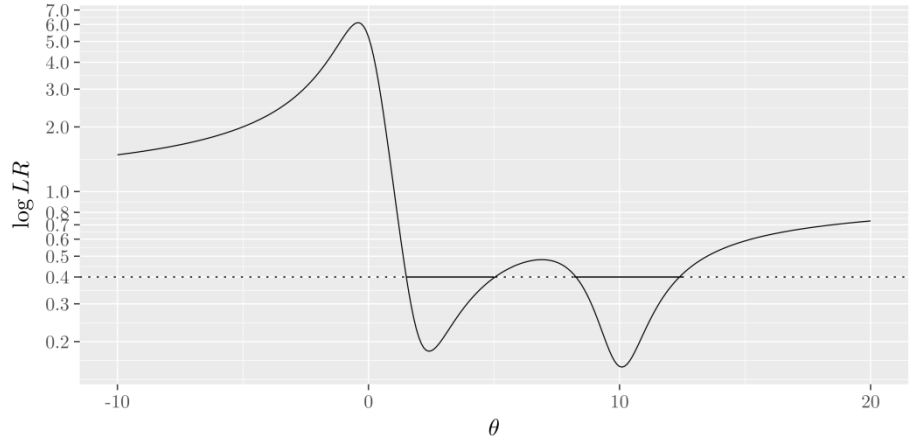


Figure 8.13 The Log Likelihood Ratio for a Mixture Alternative



8.3. Bayes' Tests

If we consider tests other than *NP* tests, it is quite clear that we can get a lower figure for stringency. All we need to do is to reduce the height of the peak, at the expense of increasing the shortcoming at other values of θ. One way to expand the class of tests is to consider a prior

probability on Θ_1, and find the most stringent Bayes' Test. If we only consider single-point priors, we get the *NP* class of tests, but as we consider other possibilities, the class of Bayes' tests is big enough to find the most stringent test.

As an illustration of the variety of new tests introduced by considering Bayes' tests, consider a two-point prior, which puts mass m at θ_1 and $1-m$ at θ. The likelihood ratio now becomes

$$\left(\frac{m}{1+(x-\theta_1)^2}+\frac{1-m}{1+(x-\theta_2)^2}\right)^{-1}\frac{1}{1+x^2}$$

To continue with the $\alpha=0.2$ example. We will let $\theta_1=\theta_{\alpha=0.2}=2\cot^{-1}0.2\pi$, $\theta_2=1$ and let the prior probability be $m=0.95$ for θ_1. The graph is Figure 8.13 shows the log likelihood ratio for this prior. We have shown the log likelihood because it emphasizes the differences in the smaller values, where the shape is unusual. As illustrated in the figure, If $C=0.4$ is chosen as the cutoff, then the rejection regions is the unions of two intervals, which is a shape not achievable with *NP* tests.

REFERENCES

Durbin, J. and G. S. Watson (1950). Testing for Serial Correlation in Least Squares Regression. I. *Biometrika*, 37 (3-4), 409–428.

Durbin, J. and G. S. Watson (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, 38 (1-2), 159–178.

Islam, T. U. (2017). Stringency-based ranking of normality tests". *Communications in Statistics - Simulation and Computation*, 46 (1), 655–668.

Khan, A. I. (2017). *Theoretical and Empirical Comparisons of CoIntegration Tests*. PhD thesis. International Islamic University of Islamabad, Pakistan.

Lehmann, E. L. and P. R. Joseph (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. New York, NY: Springer Science & Business Media, Inc. Springer e-books. ISBN: 978-0-387-27605-2.

Zaman, A. (1996). Statistical Foundations for Econometric Techniques. Academic Press.