Bahadur Efficiencies for Statistics of Truncated P-value Combination Methods

by

Xiaohui Chen

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

May 2018

APPROVED:

Professor Zheyang Wu, Major Thesis Advisor

Professor Luca Capogna, Head of Department

Abstract

Combination of p-values from multiple independent tests has been widely studied since 1930's. To find the optimal combination methods, various combiners such as Fisher's method, inverse normal transformation, maximal p-value, minimal p-value, etc. have been compared by different criteria. In this work, we focus on the criterion of Bahadur efficiency, and compare various methods under the TFisher. As a recently developed general family of combiners, TFisher cover Fisher's method, the rank truncated product method (RTP), the truncation product method (TPM, or the hard-thresholding method), soft-thresholding method, minimal p-value method, etc. Through the Bahadur asymptotics, we better understand the relative performance of these methods. In particular, through calculating the Bahadur exact slopes for the problem of detecting sparse signals, we reveal the relative advantages of truncation versus non-truncation, hard-thresholding versus soft-thresholding. As a result, the soft thresholding method is shown superior when signal strength is relatively weak and the ratio between the sample size of each p-value and the number of combining p-values is small.

Keywords: p-value combination methods, signal detection, TFisher, Bahadur efficiency.

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Zheyang Wu who made sure the thesis logically. He also gave me great support and warm encouragement when I met problems, which motivated me to find answers.

My thanks are also due to my readers who have give me useful advice in building my theories. Thanks also to Mr. Hong Zhang who also help me to get involved in this topic. I also want to thanks my parents for supporting me to study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Combination of p-values is a common practical tool for combining information across a group of hypothesis tests. In 1934, Fisher firstly presented the idea of combination of p-values with log transformation (Fisher, 1934). In 1971, Littell and Fork compared the exact slopes for fisher's method, mean of the normal transforms of the significance levels, the maximum significance level and the minimum significance level and they concluded the fisher's method enjoys the highest exact slope among these four methods (Littell and Folks, 1971). In 1973, they further proved that fisher's method is optimal among all combination methods, when finite p-values are considered and the combiner $T(T_1, ..., T_n)$ is a nondecreasing function of $T_1, ..., T_n$ (Littell and Folks, 1973).

With different assumptions and perspectives, researchers got different conclusions about optimal combination methods. Abu-Dayyeh, Al-Momani and Muttlak showed that for simple random sample (SRS) from normal distribution, the inverse normal method shares the highest exact slope as θ approaches to zero (under the alternative $H_1 : \theta > 0$). For SRS from logistics distribution, the sum of p-values has the highest exact slope as θ approaches to zero (under the alternative $H_1 : \theta > 0$)

(Abu-Dayyeh et al., 2003). M. C. Whitlock concluded that the weighted inverse normal method is superior to Fisher's combination method for normal distribution data (Whitlock, 2005). Heard proposed a rule-of-thumb for choosing p-value combination methods, based on different data sets and hypothesis tests via power (Heard and Rubin-Delanchy, 2017).

In this paper, we study a group of hypothesis test from the perspective of combination of truncated p-values, which could increase the Bahadur exact slope in some cases. We focus on log transformation and inverse normal transformation of p-values and compute the exact slopes for these transformations.

1.1 Background on Bahadur Theory

Bahadur efficiency is an important tool for choosing a efficient test statistics of large sample study. The concept of Bahaduar efficiency is firstly introduced by R. R. Bahadur in 1967, which is based on the relative rate of decreasing p-value when the sample size for each individual test goes to infinity under the alternative hypothesis. The definition of Bahadur efficiency is given here: Let the null hypothesis H_0 be $H_0 : \theta \in \Theta_0 \subset \Theta$ and the alternative H_1 be $H_1 : \theta \in \Theta_1$, where $\Theta_1 = \Theta - \Theta_0$. For any individual test statistic $T_m(x_1, x_2, ..., x_m)$, the significance level or p-value of the hypothesis test is $P_m(t) = \sup_{\theta \in \Theta_0}\{1 - F(T_m < t)\}$. If there exists a nonrandom positive function $c(\theta)$, then $c(\theta)$ is called the Bahadur exact slope or in short exact slope, such that $-\frac{2}{m} \log P_m(t) \to c(\theta), m \to \infty$ with probability one for $\theta \in \Theta_1$. The higher the exact slope is, the faster the p-values converges to zero under the alternative.

The exact slope $c(\theta)$ is positive in the original definition provided by Bahadur 1967. In theorem 1, we could further show that the exact slope $c(\theta)$ could be nonnegative:

The exact slope could be calculated by the following theorem by (Nikitin, 1995):

Theorem 1 (Bahadur)**.** *For a sequence* $\{T_m\}$*, let the following two conditions be fulfilled:*

$$T_m \to b(\theta), \theta \in \Theta_1,$$

where $-\infty < b(\theta) < \infty$*;*

$$\lim_{m\to\infty} m^{-1} \log P_m(t) = -f(t),$$

for each t *from an open interval* I *on which* f *is continuous and* $\{b(\theta), \theta \in \Theta_1\} \subset I$*. Then* $\lim_{m\to\infty} m^{-1} \log P_m = -\frac{1}{2}c(\theta)$ *is valid and, moreover, for any* $\theta \in \Theta_1$*,*

$$c(\theta) = 2f(b(\theta)).$$

Note that applying a strictly monotone increasing function $\psi(.)$ to T_m can simplify the calculation for some cases, where $T'_m = \psi(T_m)$ also satisfies these conditions in theorem 1. For example, let T_m be $T_m = X_{m_1} + X_{m_2} + ... + X_{m_n}$, where n is the number of test statistics and m_i is the sample size of each test, a strictly monotone increasing function $\psi(x) = \frac{x}{m}$ could be applied to T_m. Then for $\theta \in \Theta_1$, $\frac{T_m}{m} \to b(\theta)$. Further, for $\theta \in \Theta_0$, $\lim_{m\to\infty} m^{-1} \log [1 - F_m(T_m < mt)] = -f(t)$ and the exact slope would still be $c(\theta) = 2f(b(\theta))$.

Theorem 2. *Let* $P_m(t)$ *is the significance level for any a hypothesis test and the subscript* m *denotes the sample size for the test, for* $\theta \in \Theta_1$*, the Bahadur exact slope includes zero, i.e.* $-\frac{2}{m} \log P_m(t) \to c(\theta) = 0$ *when* $m \to \infty$*.*

Proof. The proof follows the same idea for theorem 1 in Nikitin's book page 7 (Nikitin, 1995). Assume $\lim_{m\to\infty} m^{-1} \log P_m = -f(t)$, with f being continuous on an open set I that include 0. Assume $T_m \to_P b(\theta) = 0$ under $H_1 : \theta \in \Theta_1$. Fix an arbitrary $\theta \in \Theta_1$, there exists an $\epsilon > 0$ such that $(b - \epsilon, b + \epsilon) = (-\epsilon, \epsilon) \subset I$.

Since F is monotone, $G(t) \equiv inf\{F(t;\theta) : \theta \in \Theta_0\}$ is also monotone,

$$G(-\epsilon;\theta) \leq G(T_m(s);\theta) \leq G(+\epsilon;\theta),$$

$$1 - G(+\epsilon) \leq P_m(t) \leq 1 - G(-\epsilon).$$

Taking logarithms and passing to the limit as $m \to \infty$, we obtain that:

$$-f(+\epsilon) \leq \underline{\lim}_{m\to\infty} m^{-1} \log P_m(t) \leq \overline{\lim}_{m\to\infty} m^{-1} lnP_m(t) \leq -f(-\epsilon).$$

By the continuity of f and ϵ being arbitrarily small, we obtain

$$\lim_{m\to\infty} m^{-1} \log P_m(t) = -f(0).$$

Thus, if $f(0) = 0$, we have $c(\theta) = 0$. □

1.2 Models of Hypothesis for Signal Detection Problem

Define the null hypothesis

$$H_0 : \theta \in \Theta_0 \tag{1.1}$$

and the alternative hypothesis

$$H_1 : \theta \in \Theta_1 = \Theta - \Theta_0 \tag{1.2}$$

We further specify the alternative for the signal detection problem. The first alternative is that there is only one signal in a group of hypotheses, where $i \in [1, n]$ is the index for each Bahadur exact slopes(to be studied in section 2):

$$H_1^{(1)} : c_1(\theta) > 0 \text{ and } c_i(\theta) = 0 \text{ for } i = 2, ..., n. \tag{1.3}$$

The second alternative, which is to be studied in section 4 and 5, considers the case of $k \geq 2$ signals:

$$H_1^{(2)} : c_i(\theta) > 0 \text{ for } i = 1, ..., k \text{ and } c_i(\theta) = 0 \text{ for } i = k+1, ..., n. \tag{1.4}$$

1.3 P-value Combination Methods

Let the input statistics $T_{m_1}, ..., T_{m_n}$ be independent and identically distributed random variables, where m_i is the sample size for each individual test and i is to index tests, $i \in [1, n]$. Define m be the average sample size for each individual test and n is the number of tests, i.e. $mn = m_1 + ... + m_n$. Recall that the definition of p-value of the hypothesis test is

$$P_{m_i}(t) = \sup_{\theta \in \Theta_0} \{1 - F(T_{m_i} < t)\}$$

The order statistics of the p-values are $P_{(1)} \leq ... \leq P_{(n)}$.

The general formula of a test statistic for combining these p-values is simply a multiple-to-one function of these p-values:

$$T = f(P_{m_1}, ..., P_{m_n}),$$

or equivalently a function of a monotone transformation of these p-values:

$$T = g(\bar{F}^{-1}(P_{m_1}), ..., \bar{F}^{-1}(P_{m_n})).$$

In this thesis, we consider two particular types of function $g(.)$, which are summation with potentially truncations and maximum:

$$T = \sum_{i=1}^{k} \bar{F}^{-1}(P_{m_i}) \text{ and } T = \max \bar{F}^{-1}(P_{m_i}), \text{ where } k \leq n.$$

There are different test statistics for combining these p-values:

1. Under log transformation, the Fisher's P-value combination statistics and the minimal P-value methods are:

$$T_F = \sum_{i=1}^{n} -2logP_{m_i}. \tag{1.5}$$

$$T_{Fmax} = \max(-2\log P_{m_i}). \tag{1.6}$$

2. The inverse normal transformation method (Stouffer's method):
   Let $Z_i = \bar{\Phi}^{-1}(P_i)$, the test statistic is:

$$T_N = \sum_{i=1}^{n} Z_{m_i}. \tag{1.7}$$

3. The general transformation method: Let $T_{m_i} = \bar{F}_0^{-1}(P_{m_i})$ and $m = cn$, the test statistics are:

$$T_m = \sum_{i=1}^{n} T_{m_i}. \tag{1.8}$$

$$T_{mmax} = \max T_{m_i}, \tag{1.9}$$

   where $F(.)$ satisfies $1 - F(\sum_{i=1}^{n} T_{m_i} < \sqrt{m}t) = O(m^c(1 - (F(T_{m_i} < \sqrt{m}t))^n))$.

4. The rank truncated product method (RTP) (Dudbridge and Koeleman, 2003):

$$W_R = \prod_{i=1}^{k^*} P_{(i)}, \text{ for } 1 \leq k^* \leq n. \tag{1.10}$$

Applying a monotone transformation, we obtain the test statistic T_R:

$$T_R = -2logW_R = \sum_{i=1}^{k^*} -2logP_{(i)}. \tag{1.11}$$

5. We also consider a recent family of statistics called "TFisher" T_S, which is analogous to RTP formula (Zhang et al., 2018):

$$W_S = \prod_{i=1}^{n} \left(\frac{P_{m_i}}{\tau_2}\right)^{I(P_{m_i} \leq \tau_1)}. \tag{1.12}$$

When $\tau_1 = \tau_2 = \tau$, it becomes the soft-thresholding:

$$W_s = \prod_{i=1}^{n} \left(\frac{P_{m_i}}{\tau}\right)^{I(P_{m_i} \leq \tau)}. \tag{1.13}$$

When $\tau_1 = \tau$ and $\tau_2 = 1$, the test statistic called the hard-thresholding:

$$W_h = \prod_{i=1}^{n} {P_{m_i}}^{I(P_{m_i} \leq \tau)}, \tag{1.14}$$

which is also called the TPM statistic Dudbridge and Koeleman (2003).

Chapter 2

Exact Slopes for One Signal

In this chapter, we study the exact slopes for log transformation and inverse normal transformation of p-value combination for the case of one signal defined in (1.3):

2.1 The Exact Slope for Log Transformation

We firstly introduce a lemma for deducing the exact slope of fisher's log transformed statistic:

Lemma 1. *Let* $t > 0$, $n = o(m)$ *and* $m \to \infty$. *Then, for the sequence* $x_i = \frac{(mt/2)^{i-1}}{(i-1)!}$, $i \in [1, n]$, *we have*

$$x_n \gg \sum_{i=1}^{n-1} x_i.$$

Proof. Consider the ratio of ith to $(i-1)$th term in this sequence is $\frac{x_i}{x_{i-1}} = \frac{mt/2}{i-1}$. Since i is the index from 1 to n such that $n = o(m)$, the ratio of two consecutive terms goes to infinity, as $m \to \infty$, i.e. $x_{i-1} = o(x_i)$. Similarly, we have $x_{i-2} = o(o(x_i))$. In this case, the summation of first $n-1$th term $\sum_{i=0}^{n-1} x_i = o(x_n)$. So, $x_n \gg \sum\limits_{i=1}^{n-1} x_i$. □

Intuitively speaking, when a series includes the ratio of power function of a value approaching infinity to a factorial of a positive integer, the nth item could represent the summation of this series, since the summation of terms from the first to $(n-1)$th is dominated.

However, when the increasing rate of n and m are same, i.e. $m = cn$, the Lemma 1 may not be true. The reasons are as follow:

For the case of $m = cn$, the ratio of nth term and $(n-1)$th term is a constant,

$$\frac{x_n}{x_{n-1}} = \frac{(nct/2)^{n-1}}{(n-1)!} \Big/ \frac{(nct/2)^{n-2}}{(n-2)!} = \frac{ct}{2}, \text{ as } n, m \to \infty$$

while the ratio of 2^{nd} and 1^{nd} term goes to infinty,

$$\frac{x_2}{x_1} = \frac{(nct/2)^1}{1!} \Big/ 1 = \frac{nct}{2} \to \infty, \text{ as } n, m \to \infty.$$

Thus, the ratio of any term to previous term is decreasing, and when $n, m \to \infty$ the ratio of two consecutive terms is a constant. Therefore, Lemma 1 is not valid when $m = cn$. However, we could still approximate the value of $\log \sum_{i=1}^{n} x_i$ by x_n, as $n, m \to \infty$. The following lemma says that although the summation of first $(n-1)$th terms could not be dominated when $m = cn$, after transformation of logarithm, the summation could still be represented by the last term.

Lemma 2. *Let $m = cn \to \infty$ and $t > 0$. Then for the sequence $x_i = \frac{(mt/2)^{i-1}}{(i-1)!}$, $i \in [1, n]$, we have*

1.

$$\log \sum_{i=1}^{n} x_i \sim \log x_n.$$

2.

$$-\frac{1}{m}\log\sum_{i=1}^{n}x_i \to -\frac{1}{c}\Big[\log(ct/2)+1\Big], \text{ as } n\to\infty.$$

Proof. Because the ratio of any term to the previous term is not less than the constant value $ct/2$. We have

$$\log x_n \le \log\sum_{i=1}^{n}x_i \le \log nx_n,$$

$$\lim_{n\to\infty}\frac{\log nx_n}{\log x_n} = \lim_{n\to\infty}\frac{\log x_n+\log n}{\log x_n} = 1+\lim_{n\to\infty}\frac{\log n}{\log x_n}. \tag{2.1}$$

Further, by L'Hopital rule, we obtain

$$\lim_{n\to\infty}\frac{\log n}{\log x_n} = \lim_{n\to\infty}\frac{1}{n\log' x_n},$$

where

$$\log' x_n = \frac{d}{dn}\log\frac{(cnt/2)^{n-1}}{(n-1)!} = \frac{((cnt/2)^{n-1})'}{(cnt/2)^{n-1}} - \frac{(n-1)!'}{(n-1)!}. \tag{2.2}$$

The first term in the right hand side of equation (2.2) is

$$\frac{((cnt/2)^{n-1})'}{(cnt/2)^{n-1}} = \frac{(cnt/2)^{n-1}(\log cnt/2+1)}{(cnt/2)^{n-1}} = \log(cnt/2)+1.$$

By stirling's approximation, i.e. $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$, the second term in the right hand side of equation (2.2) is

$$\begin{aligned}\frac{(n-1)!'}{(n-1)!} &= \frac{\sqrt{2\pi}e^{1-n}(n-1)^{n-\frac{1}{2}}\log(n-1)+\frac{n-\frac{1}{2}}{n-1}-1}{\sqrt{2\pi}e^{1-n}(n-1)^{n-\frac{1}{2}}}\\ &= \log(n-1)+\frac{n-\frac{1}{2}}{n-1}-1 = \log(n-1)\to\infty, \text{ as } n\to\infty.\end{aligned}$$

Then, follow the equation (2.2), we have

$$\lim_{n\to\infty} \log' x_n = \lim_{n\to\infty} \log{(cnt/2)}/(n-1) + 1 = \log{(ct/2)} + 1. \tag{2.3}$$

Continue with equation (2.1),

$$\lim_{n\to\infty} \frac{\log nx_n}{\log x_n} = 1.$$

So,

$$\log \sum_{i=1}^{n} x_i \sim \log x_n.$$

Thus, the logarithm of summation could asymptotically equal to the logarithm of the biggest term in this series.

To prove the second part in this lemma, apply the result of the first part in the lemma and equation (2.3),

$$-\frac{1}{m} \log \sum_{i=1}^{n} x_i \sim -\frac{1}{m} \log x_n = -\frac{1}{cn} \log \frac{(cnt/2)^{n-1}}{(n-1)!} = -\frac{1}{c}\Big[\log{(ct/2)} + 1\Big],$$

as $n \to \infty$.

□

Next, we provide the Bahadur exact slope of fisher's log transformation method:

Theorem 3. *Under the alternative (1.3), the exact slope of the Fisher's P-value combination statistic (1.5) is*

$$c_F(\theta) = \begin{cases} c_1(\theta), \text{ when } n \text{ is finite and } m \to \infty \\ c_1(\theta) - \frac{2}{c}\Big[\log{(c_1(\theta)c/2)} + 1\Big], \text{ when } m = cn \to \infty. \end{cases}$$

Proof. Under the alternative (1.3),

$$\frac{T_F}{m} = \sum_{i=1}^{n} \frac{-2\log P_{m_i}}{m} \to c_1(\theta).$$

Note that under H_0, T_F follows chi-square distribution with 2n degrees of freedom and its cumulative distribution function is $F_F(x) = \frac{\gamma(2n/2, x/2)}{\Gamma(2n/2)}$, where $\gamma(.)$ is the lower incomplete gamma function and a general series representation is $\gamma(n,z) = (n-1)!(1-e^{-z}(\sum_{i=0}^{n-1}\frac{z^i}{i!}))$(Koziol and Tuckwell, 1999).
When n is finite, $n = o(m)$, under the null hypothesis,

$$\begin{aligned}
-\frac{1}{m}\log\left[1-F_F(mt)\right] &= -\frac{1}{m}\log\left[e^{-mt/2}(\sum_{i=0}^{n-1}\frac{(mt/2)^i}{i!})\right], m \to \infty \\
&= -\frac{1}{m} \times (\frac{-mt}{2}) - \frac{1}{m}\log\left[\frac{(mt/2)^{n-1}}{(n-1)!} + \sum_{i=0}^{n-2}\frac{(mt/2)^i}{i!}\right]
\end{aligned}$$

By lemma 1,

$$-\frac{1}{m}\log\left[1-F_F(mt)\right] \sim \frac{1}{2}t - \frac{1}{m}\log\frac{(mt/2)^{n-1}}{(n-1)!}, m \to \infty$$

By L'Hospital's rule,

$$\lim_{m\to\infty} -\frac{1}{m}\log\left[1-F_F(mt)\right] = \lim_{m\to\infty}\frac{1}{2}t - \frac{n-1}{m} = \frac{1}{2}t.$$

To guarantee the right hand side $\frac{1}{2}c_i(\theta) - \frac{n-1}{m} \geq 0$, n cannot be too big. That is, $n \leq \frac{c_1(\theta)m}{2} + 1$ when this condition is satisfied by Theorem 1. The exact slope of (1.5) is

$$c_F(\theta) = c_1(\theta).$$

When $m = cn$, by lemma 2,

$$\lim_{m\to\infty} -\frac{1}{m}\log\left[1 - F_F(mt)\right] = \frac{1}{2}t - \frac{1}{c}\Big[\log\left(ct/2\right) + 1\Big],$$

Thus, by Theorem 1

$$c_F(\theta) = c_1(\theta) - \frac{2}{c}\Big[\log\left(c_1(\theta)c/2\right) + 1\Big].$$

□

Then, consider the case of making a log transformation of the minimum p-value or equivalently the maximum $-2\log P_{m_i}$:

Theorem 4. *Under the alternative (1.3), the exact slope of minimal P-value statistic (1.6) is*

$$c_{Fmax}(\theta) = c_1(\theta) \text{ when } n \leq cm \text{ and } c \text{ is a constant.}$$

Proof. For the maximum of random variables:

$$T_{Fmax} = -2\log\min P_{m_i} = \max(-2\log P_{m_i})$$

Under the alternative,

$$\frac{T_{Fmax}}{m} \to c_1(\theta)$$

Under the null hypothesis, $-2\log P_{m_i}$ follows chi-square distribution with 2 degrees of freedom, which is exponential with parameter $\lambda = \frac{1}{2}$.

$$\begin{aligned}
1 - F_{Fmax}(mt) &= 1 - P(max(-2\log P_{m_i}) < mt) \\
&= 1 - (P(-2\log P_{m_i} < mt))^n \\
&= 1 - (1 - e^{-\frac{1}{2}mt})^n \\
&= 1 - (1 - ne^{-\frac{1}{2}mt} + o(e^{-mt})) \\
&= ne^{-\frac{1}{2}mt}
\end{aligned}$$

and

$$-\frac{1}{m}\log ne^{-\frac{1}{2}mt} = -\frac{1}{m}(\log n - \frac{1}{2}mt) = \frac{1}{2}t - \frac{\log n}{m}, m \to \infty.$$

To guarantee the right hand side $\frac{1}{2}c_i(\theta) - \frac{\log n}{m}$, n cannot be too big. That is, $n \leq e^{\frac{1}{2}c_1(\theta)m}$ (note we replace t by $c_1(\theta)$). As $-\frac{\log n}{m} \to 0$, the exact slope is

$$c_{Fmax}(\theta) = c_1(\theta) - \frac{2logn}{m} = c_1(\theta), \text{when } n \leq cm \text{ and } c \text{ is a constant.}$$

□

The maximum and summation perform equally, since they share the same exact slope when the number of hypothesis tests n is finite. However, when $m = cn \to \infty$, the truncated method with fisher's log-transformation could have a larger exact slope than the non-truncated method when $c_1(\theta)c \geq 2\log(\frac{c_1(\theta)c}{2}) + 2$.

2.2 The Exact Slopes for Inverse Normal Transformation

Besides log transformation, inverse normal transformation is also commonly used in practice. In this section, our purpose is to get the exact slopes from inverse normal transformation with and without truncation.

Theorem 5. *Under the alternative (1.3), the exact slope of (1.7) is* $c_N = \frac{c_1(\theta)}{n}$.

Proof. Assume $m_i = m, i = 1, ..., n$. Since $\frac{1}{m}[Z_{m_i}]^2 \to c_i(\theta)$ with probability one (Littell and Folks, 1971), we have

$$\frac{T_N}{\sqrt{m}} \to \sum_{i=0}^{n} \sqrt{c_i(\theta)} = \sqrt{c_1(\theta)},$$

where $c_1(\theta) > 0$ and $c_i(\theta) = 0$ for $i = 2, ..., n$.

Under H_0,

$$1 - F_N(\sqrt{m}t) = 1 - P(T_N < \sqrt{m}t) = 1 - P(\frac{T_N}{\sqrt{n}} < \frac{\sqrt{m}t}{\sqrt{n}}) = \bar{\Phi}(\frac{\sqrt{m}t}{\sqrt{n}})$$

By Mill's ratio,

$$\bar{\Phi}(\frac{\sqrt{m}t}{\sqrt{n}}) \sim \frac{\phi(\frac{\sqrt{m}t}{\sqrt{n}})}{\frac{\sqrt{m}t}{\sqrt{n}}} = \frac{\sqrt{n}}{\sqrt{2\pi m}t} e^{-\frac{mt^2}{2n}} \sim e^{-\frac{mt^2}{2n}}, \text{as } m \to \infty \text{ and } n = o(m).$$

So,

$$-\frac{2}{m} log e^{-\frac{mt^2}{2n}} = \frac{t^2}{n}.$$

Thus, $c_m^{(N)} = \frac{c_1(\theta)}{n}$. When $n \to \infty$, $c_m^{(N)} = 0$. □

Theorem 6. *Under the alternative (1.3), the exact slope for maximum is* $c_{Nmax} = c_1(\theta)$.

Proof. Assume $m_i = m, i = 1, ..., n$. Under alternative (1.3), we have

$$\frac{T_{Nmax}}{\sqrt{m}} = \frac{max Z_{m_i}}{\sqrt{m}} \to max(\sqrt{c_i(\theta)}) = \sqrt{c_1(\theta)}$$

Under H_0, by Mill's ratio,

$$1 - F_{Nmax}(\sqrt{m}t) = 1 - P(T_{Nmax} < \sqrt{m}t) = 1 - (P(Z_{m_i} < \sqrt{m}t))^n$$

$$= 1 - (1 - \bar{\Phi}(\sqrt{m}t))^n = 1 - (1 - \frac{1}{\sqrt{2\pi}}e^{-\frac{mt^2}{2}}\frac{1}{\sqrt{m}t})^n = \frac{n}{\sqrt{2\pi m}t}e^{-\frac{mt^2}{2}} \sim e^{-\frac{mt^2}{2}}.$$

So,

$$-\frac{2}{m}loge^{-\frac{mt^2}{2}} = t^2 = c_1(\theta).$$

Thus, $c_{Nmax} = c_1(\theta)$, for any $n \leq cm$. □

Based on Theorem 5 and 6, we have the following conclusion: The maximum and sum of p-values with inverse normal transformation do not share the same tail distribution. Moreover, the the exact slope of maximum Z_i, where i is from 1 to n, is higher than the one of summation. Thus, the truncated normal distribution method has a higher slope than the original non-truncated normal-transformation method.

2.3 More General Transformations

Here we give a sufficient condition such that summation based statistic has the same Bahardur slope as the maximum based statistic, when the number of tests n is finite.

We consider a more general type of transformation $\bar{F}^{-1}()$, where F is a cumulative density function and $\bar{F} = 1 - F$ is the survival funcation. T_{m_i} is defined by $T_{m_i} = \bar{F}^{-1}(P_{m_i})$. The summation based statistic:

$$T_m = T_{m_1} + ... + T_{m_n}$$

The maximum based statistics:

$$T_{max} = \max T_{m_i}$$

Under the alternative (1.3), as with the case of summation, we have

Assume under the alternative (1.3),

$$\frac{T_m}{m} = \frac{T_{max}}{m} \to b(\theta), \theta \in \Theta_1$$

Under H_0,

$$-\frac{1}{m} log(1 - F(T_m < mt)) \to f(t), \tag{2.4}$$

and

$$-\frac{1}{m} log(1 - F(T_{m_i} < mt)^n) \to f(t) \tag{2.5}$$

If $F(\sum T_{m_i} < mt) \sim (F(T_{m_i} < mt))^n$ i.e., T_{m_i} follows a so-called subexponential distribution (Pitman, 1980; Goldie and Klüppelberg, 1998), the maximum and the sum based on statistics share the same right-tail rate. Then, the maximum and the sum based statistics also share the same exact slope.

Further, because we are comparing the ratio of the log tail probability, we still get the same slope if there exists a constant c such that $1 - F(\sum T_{m_i} < mt) = O(m^c(1 - (F(T_{m_i} < mt))^n))$ holds.

Due to the fact that the family of statistics with fisher's log transformation follow Chi-square distribution, which is not such a subexponential distribution(The proof for this statement is in Appendix A.1.), but satisfies $1 - F(\sum T_{m_i} < mt) = O(m^c(1 - (F(T_{m_i} < mt))^n))$, and thus has the same slope. In this case, generalize the transformation $\bar{F}^{-1}(.)$ to inverse exponential and inverse gamma transformation (under traditional definition).

Theorem 7. *Under the alternative (1.3), if the cumulative distribution function* *satisfies*

$$1 - F(\sum T_{m_i} < \sqrt{m}t) = O(m^c(1 - (F(T_{m_i} < \sqrt{m}t))^n)), \tag{2.6}$$

the maximum and summation of T_{m_i} share the same exact slope, where c is a constant.

Proof. Continue the previous results (2.4) and (2.5). When $1 - F(\sum T_{m_i} < \sqrt{m}t) = O(m^c(1 - (F(T_{m_i} < \sqrt{m}t))^n))$ holds, and $m \to \infty$, we obtain

$$-\frac{1}{m}\log(1 - F(\sum T_{m_i})) \sim -\frac{1}{m}\log(m^c(1 - F(T_{m_i}))^n) \to f(t).$$

That is under the null, $-\frac{1}{m}\log(1 - F(\sum T_{m_i})) \sim -\frac{1}{m}\log(1 - F(\max T_{m_i}))$. Also, under the alternative (1.3), the maximum and summation based statistics converge to the same value. Overall, the maximum and summation of T_{m_i} share the same exact slope. $\square$

Corollary 1. *When the number of test statistics n is finite, summation and maximum with inverse exponential transformation or inverse gamma transformation of significance levels share the same exact slope under the alternative (1.3).*

Proof. Assume T_{m_i} follows exponential distribution with parameter λ under H_0, the test statistics (1.8) follows $Gamma(n, \lambda)$. Then, the probability of $\max T_{m_i}$ and $\sum T_{m_i}$ are:

$$(P(T_{m_i} < mt))^n = (1 - e^{-\lambda mt})^n \sim 1 - ne^{-\lambda mt},$$

and

$$P(\sum T_{m_i} < mt) = 1 - e^{-\lambda mt} \sum_{i=0}^{n-1} \frac{(\lambda mt)^i}{i!}.$$

The value of $\frac{(\lambda mt)^i}{i!}$ largely increases as n increases and $m \to \infty$. Similarly, $\sum_{i=0}^{n-1} \frac{(\lambda mt)^i}{i!} \sim \frac{(\lambda mt)^{n-1}}{(n-1)!}$.

$$P(\sum T_{m_i} < mt) \sim 1 - e^{-\lambda mt}\frac{(\lambda mt)^{n-1}}{(n-1)!}$$

Thus, the maximum and summation with inverse exponential transformation satisfies the formula (2.6).

Next, assume T_{m_i} follows a gamma distribution $Gamma(\alpha, \lambda)$, the test statistic (1.8) follows a $Gamma(n\alpha, \lambda)$. Then, the probability of $maxT_{m_i}$ and $\sum T_{m_i}$ are:

$$\begin{aligned}
(P(T_{m_i} < mt))^n &= [\frac{(\alpha-1)!(1 - e^{-\lambda mt}\sum_{i=0}^{\alpha-1}\frac{(\lambda mt)^i}{i!})}{\Gamma(\alpha)}]^n \\
&= (1 - e^{-\lambda mt}\sum_{i=0}^{\alpha-1}\frac{(\lambda mt)^i}{i!})^n \\
&\sim 1 - ne^{-\lambda mt}\sum_{i=0}^{\alpha-1}\frac{(\lambda mt)^i}{i!} \\
&\sim 1 - ne^{-\lambda mt}\frac{(\lambda mt)^{\alpha-1}}{(\alpha-1)!},
\end{aligned}$$

and

$$\begin{aligned}
P(\sum T_{m_i} < \sqrt{m}t) &= \frac{(n\alpha-1)!(1 - e^{-\lambda mt}\sum_{i=0}^{n\alpha-1}\frac{(\lambda mt)^i}{i!})}{\Gamma(n\alpha)} \\
&= 1 - e^{-\lambda mt}\sum_{i=0}^{n\alpha-1}\frac{(\lambda mt)^i}{i!} \\
&\sim 1 - e^{-\lambda mt}\frac{(\lambda mt)^{n\alpha-1}}{(n\alpha-1)!},
\end{aligned}$$

Thus, the summation and maximum with inverse gamma transformation also satisfies the formula (2.6). We can conclude that the summation and maximum with inverse exponential and gamma transformation share the same exact slope for one signal case. $\square$

Chapter 3

Exact Slopes for k Signals

In this chapter, we further extend the the number of signals from one to k. The Bahadur exact slopes of RTP and TFisher are given in the section 3.1 and section 3.2.

3.1 Rank Truncated Product(RTP)

In this section, we study the Bahadur slope for rank truncated product based on Fisher's log transformation.

Following the result of RTP test (Dudbridge and Koeleman, 2003), the exact distribution of W_R in (1.10) is

$$P(W_R \leq w) = \binom{n}{k^*+1}(k^*+1)\int_{v_0}^{1}(1-v)^{n-k^*-1}A(w,v)dv, \tag{3.1}$$

where

$$A(w,v_0) = \begin{cases} w\sum_{s=0}^{k^*-1}\frac{(k^*lnv_0-lnw)^s}{s!}, & \text{when } w \leq v_0^{k^*} \\ v_0^{k^*}, & \text{otherwise} \end{cases} \tag{3.2}$$

After applying log-transformation, $T_R = -2logW_R = \sum_{i=1}^{k^*} -2logP_{(i)}$, we obtain the cumulative density function of T_R :

$$P(T_R \geq mt) = P(-2logW_R \geq mt) = P(logW_R \leq -\frac{m}{2}t) = P(W_R \leq e^{-\frac{m}{2}t}),$$

$$P(T_R \geq mt) = \binom{n}{k^*+1}(k^*+1)e^{-\frac{m}{2}t}\int_{v_0}^{1}(1-v)^{n-k^*-1}\sum_{s=0}^{k^*-1}\frac{(k^*\log v - \log e^{-\frac{m}{2}t})^s}{s!}dv, \tag{3.3}$$

when $w \leq v_0^{k^*}$.

Now, we intend to derive the exact slope of RTP. Before the deduction, a useful lemma used in the proof of exact slope of RTP is introduced as following:

Lemma 3. *By mathematical induction,* $\frac{(\frac{m}{2}t+k^*lnv)^{(k^*-1)}}{(k^*-1)!} \gg \sum_{s=0}^{k^*-2}\frac{(\frac{m}{2}t+k^*lnv)^s}{s!}$, *for constant* $k^* \geq 2$, $v \in (0,1)$ *and* $m \to \infty$.

Proof. When $s = 1$, $k^*lnv+\frac{m}{2}t \gg 1$. Show that if $\frac{(k^*lnv+\frac{m}{2}t)^{k^*-2}}{(k^*-2)!} \gg \sum_{s=0}^{k^*-3}\frac{(k^*lnv+\frac{m}{2}t)^s}{s!}$ holds, $\frac{(k^*lnv+\frac{m}{2}t)^{(k^*-1)}}{(k^*-1)!} \gg \sum_{s=0}^{k^*-2}\frac{(k^*lnv+\frac{m}{2}t)^s}{s!}$ holds.

$$\begin{aligned}
k^*lnv + \frac{m}{2}t &\gg 2k^* - 2 \\
\frac{(k^*lnv + \frac{m}{2}t)^{k^*-1}}{(k^*-1)!} &\gg 2\frac{(k^*lnv + \frac{m}{2}t)^{k^*-2}}{(k^*-2)!} \\
\frac{(k^*lnv + \frac{m}{2}t)^{(k^*-1)}}{(k^*-1)!} &\gg \frac{(k^*lnv + \frac{m}{2}t)^{k^*-2}}{(k^*-2)!} + \sum_{s=0}^{k^*-3}\frac{(k^*lnv + \frac{m}{2}t)^s}{s!}
\end{aligned}$$

□

Then, we provide the Bahadur exact slope for RTP:

Theorem 8. *Under the alternative hypothesis (1.4), the exact slope of RTP based on Fisher's log transformation (1.11) is* $\sum_{i=1}^{k\wedge k^*} c_i(\theta)$, *where* k^* *is a constant in* $[1, n]$.

Proof. Under the alternative hypothesis,

When $k^* > k$,

$$\frac{T_R}{m} \to \sum_{i=1}^{k} c_i(\theta).$$

Since the smallest noise p-value is bigger than $-2\log U_{(1)}$, where $U_{(1)} = \min_{1\leq i\leq n-k^*} U_i$. $-2\log U_{(1)} \overset{D}{=} \max\{X_i^2 + Y_i^2, i = 1, ..., n\} \leq X_{(n)}^2 + Y_{(n)}^2 \sim 2(\sqrt{2\log n})^2 \ll m$, where X_i and Y_i are iid $N(0,1)$. The exact slope of the smallest noise p-value is $-\frac{2}{m}\log U_{(1)} \leq \frac{1}{m}(X_{(n)}^2 + Y_{(n)}^2) \sim \frac{2}{m}(\sqrt{2\log n})^2 \to 0$. Moreover, since the smallest noise p-value has zero exact slope, the $K-k$ smallest noise p-values are all zero exact slope. Compared with the exact slopes of signal p-values, the ones of noise p-values could be ignored.

When $k^* \leq k$,

$$\frac{T_R}{m} \to \sum_{i=1}^{k^*} c_i(\theta).$$

Under the null hypothesis, when $m \to \infty$ and $w = e^{-mt/2} \leq v_0^{k^*}$, the another case is given latter in the note of this proof.

$$\begin{aligned}
&-\frac{1}{m}log(1-F(mt))\\
&= -\frac{1}{m}logP(W_R \leq e^{-\frac{m}{2}t})\\
&= -\frac{1}{m}log[\binom{n}{k^*+1}(k^*+1)\int_{v_0}^{1}(1-v)^{n-k^*-1}e^{-\frac{m}{2}t}\sum_{s=0}^{k^*-1}\frac{(k^*lnv - lne^{-\frac{m}{2}t})^s}{s!}dv]\\
&= -\frac{1}{m}log[\binom{n}{k^*+1}(k^*+1)e^{-\frac{m}{2}t}\int_{v_0}^{1}(1-v)^{n-k^*-1}\sum_{s=0}^{k^*-1}\frac{(k^*lnv + \frac{m}{2}t)^s}{s!}dv]\\
&= -\frac{1}{m}[log\binom{n}{k^*+1}(k^*+1) - \frac{m}{2}t + log\int_{v_0}^{1}(1-v)^{n-k^*-1}\sum_{s=0}^{k^*-1}\frac{(k^*lnv + \frac{m}{2}t)^s}{s!}dv]\\
&= \frac{1}{2}t - \frac{1}{m}log\int_{v_0}^{1}(1-v)^{n-k^*-1}\sum_{s=0}^{k^*-1}\frac{(k^*lnv + \frac{m}{2}t)^s}{s!}dv
\end{aligned}$$

According by lemma 3,

$$
\begin{aligned}
&-\frac{1}{m}log(1-F(mt)) \\
&\sim \frac{1}{2}t-\frac{1}{m}log\frac{1}{(k^*-1)!}-\frac{1}{m}log\int_v^1(1-v)^{n-k^*-1}(k^*lnv+\frac{m}{2}t)^{(k^*-1)}dv \\
&\sim \frac{1}{2}t-\frac{1}{m}log\int_{v_0}^1(1-v)^{n-k^*-1}(k^*lnv+\frac{m}{2}t)^{k^*-1}dv
\end{aligned}
$$

Since $(1-v)^{n-k^*-1}(k^*lnv+\frac{m}{2}t)^{k^*-1}\sim(\frac{m}{2}t)^{k^*-1}(1-v)^{n-k^*-1}$,

$$
\begin{aligned}
\int_{v_0}^1(1-v)^{n-k^*-1}(k^*lnv+\frac{m}{2}t)^{k^*-1}dv &\sim \int_{v_0}^1(\frac{m}{2}t)^{k^*-1}(1-v)^{n-k^*-1}dv \\
&\sim -\frac{(\frac{m}{2}t)^{k^*-1}}{n-k^*}(1-v)^{n-k^*}\Big|_v^1 \\
&\sim \frac{(\frac{m}{2}t)^{k^*-1}}{n-k^*}(1-v)^{n-k^*}
\end{aligned}
$$

Further, $\lim_{m\to\infty}\frac{1}{m}log\frac{(\frac{m}{2}t)^{k^*-1}}{n-k^*}(1-v)^{n-k^*}=\frac{k^*-1}{m}\to 0$.

We have $-\frac{2}{m}log(1-F(mt))\sim t$. Thus, the exact slope of RTP based on Fisher's log transformation is $\sum_{i=1}^{k\wedge k^*}c_i(\theta)$. Note that when $w\geq v_0^{k^*}$ and k^* is a constant, $-\frac{1}{m}\log(1-F(mt))=-\frac{1}{m}\log\binom{n}{k^*+1}(k^*+1)\int_{v_0}^1(1-v)^{n-k^*-1}v^{k^*}dv=0$. In this case, the exact slope is 0.

$\square$

The exact slope of rank truncated product is depended on the choice of truncation k^*. If k^* is less than the number of nonzero signals, the exact slope is the summation of $c_i(\theta)$, where $i=1,...,k^*$; otherwise, the exact slope is the summation of $c_i(\theta)$, where $i=1,...,k$.

3.2 TFisher

In this section, we study a more general test statistic with weight and truncation called "TFisher" in (1.12) under the alternative hypothesis (1.4). We derive the lower bounds and upper bounds Bahadur exact slope for "TFisher", based on the relationship of the number of tests n and the sample size m. Then, we compare the Bahadur exact slopes from different combination of τ_1 and τ_2.

Here, the test statistic we considered is:

$$W_S = \prod_{i=1}^{n} \left(\frac{P_i}{\tau_2}\right)^{I(P_i \leq \tau_1)}.$$

Taking a logarithm of W_S, we have

$$T_S = 2K log\tau_2 - 2\sum_{i=1}^{K} \log P_{(i)}, \tag{3.4}$$

where random variable $K = \#\{P_i \leq \tau_1\}$. Under the null, $K \sim Binomial(n, \tau_1)$, so the mean is $E(K) = n\tau_1 \equiv k$. Based on deductions in Zhang et al. (2018) for p-value calculation, since the density of W is derived from Chi-square distribution, when $t_0 + 2k\log(\tau_1/\tau_2) \geq 0$, the density of W is

$$P(T_S \geq t_0) = (1-\tau_1)^n I_{\{t_0 \leq 0\}} + e^{-t_0/2} \sum_{k=1}^{n}\sum_{j=0}^{k-1} \binom{n}{k} \tau_2^k (1-\tau_1)^{n-k} \frac{(t_0 + 2k\log(\tau_1/\tau_2))^j}{2^j j!}; \tag{3.5}$$

3.2.1 Convergence of T_S/m for TFisher

In this subsection, we provide the convergency in probability for T_S/m. Two parts are showed as follow: first, study the convergence of the first term in T_S/m, $\frac{2K}{m} log\tau_2$, which is provided in Lemma 4 and then study the convergence of $-\frac{2}{m}\sum_{i=1}^{K} logP_{(i)}$

provided in Lemma 5.

Lemma 4. *Let k be the number of signals and K be the number of P_i, where $P_i \leq \tau_1$ under H_0. If the ratio of the number of tests and the average sample size for each individual test , $\frac{n}{m}$, converges to zero as $m \to \infty$, i.e. $n = o(m)$, then $\frac{2K}{m}\log\tau_2 \to 0$; If $\frac{n}{m}$ converges to a constant c as $m \to \infty$, i.e. $m = cn$, then $\frac{2K}{m}\log\tau_2 \to \frac{2\tau_1}{c}\log\tau_2$.*

Proof. When $n = cm$ as $m \to \infty$, by Chebyshev's inequality, for any $\epsilon > 0$,

$$P(|\frac{K}{n} - \tau_1| \geq \epsilon) \leq \frac{\tau_1(1-\tau_1)}{n\epsilon^2} \to 0, n \to \infty$$

Thus, we have $K/n \to \tau_1$. Then, we have $\frac{2K}{m}log\tau_2 \to \frac{2\tau_1}{c}log\tau_2$. When $\frac{n}{m} \to 0$, $\frac{2K}{m}log\tau_2 \to 0$. □

Lemma 5. *If the ratio of the average sample size for each individual test and the number of tests, $\frac{n}{m}$, converges to zero as $m \to \infty$, i.e. $n = o(m)$, then:*

a. $P(K < k) \to 0$.

b. $-\frac{2}{m}\sum_{i=1}^{K} logP_{(i)} \to \sum_{i=1}^{k} c_i$, when $K \geq k$.

If the ratio of the average sample size for each individual test and the number of tests, $\frac{n}{m}$, converges to a constant c as $m \to \infty$, then:

a. $P(K < k) \to 0$.

b. $-\frac{2}{m}\sum_{i=1}^{K} logP_{(i)} \to \sum_{i=1}^{k} c_i + \frac{2\tau_1}{c} + C$, when $K \geq k$ and fixed $C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c})$.

Proof. First, consider the case of $K < k$ and we can show that $P(K < k) \to 0$. Assume $P_i, i = 1, .., k$, are nonzero signals, for any ϵ, such that $|P_i - 0| < \epsilon, i = 1, ..., k$. Besides, K is the number of $P_i < \tau_1$, where τ_1 is a nonzero fixed number. If $K < k$, we obtain $|P_i - 0| < \tau_1 < \epsilon$. Yet, a nonzero constant could not less than an arbitrary ϵ. Thus, $P(K < k) \to 0$.

Furthermore, consider the case of $K = k$, which means the number of $P_i < \tau_1$ is the same as the number of nonnegative signals. We can show that

$$-\frac{2}{m}\sum_{i=1}^{k} logP_{(i)} = -\frac{2}{m}\sum_{i=1}^{k} logP_{(i)} \rightarrow \sum_{i=1}^{k} c_i(\theta), \text{ as } m \rightarrow \infty. \tag{3.6}$$

Let X be $-\frac{2}{m}\sum_{i=1}^{k} logP_{(i)}$, Y be $-\frac{2}{m}\sum_{i=1}^{k} logP_i$, Z be $\sum_{i=1}^{k} c_i(\theta)$ and A be $\{X = Y\}$.

For any $\epsilon > 0$,

$$\begin{aligned} P(|-\frac{2}{m}\sum_{i=1}^{k} logP_{(i)} + \frac{2}{m}\sum_{i=1}^{k} logP_i| > \epsilon) &= P((-\frac{2}{m}\sum_{i=1}^{k} logP_{(i)} + \frac{2}{m}\sum_{i=1}^{k} logP_i)^2 > \epsilon^2) \\ &\leq \frac{4E(-\sum_{i=1}^{k} logP_{(i)} + \sum_{i=1}^{k} logP_i)^2}{m^2\epsilon^2} \end{aligned}$$

Since $-\frac{2}{m} logP_i \rightarrow c_i(\theta) > 0$, $P_i \rightarrow 0$, $m \rightarrow \infty$ for $i = 1, ..., k$. Then, the first k ordered p-values have $\{P_{(i)}, i = 1, ...k\} = \{P_i, i = 1, ...k\}$ and $\sum_{i=1}^{k} logP_{(i)} = \sum_{i=1}^{k} logP_i$. The expectation of square of difference of X and Y convergence to 0. So,

$$P(|-\frac{2}{m}\sum_{i=1}^{k} logP_{(i)} + \frac{2}{m}\sum_{i=1}^{k} logP_i| > \epsilon) \rightarrow 0, i.e. P(A) \rightarrow 1.$$

Next,

$$\begin{aligned} P(|X - Z| \geq \epsilon) &= P(|X - Z| \geq \epsilon | A)P(A) + P(|X - Z| \geq \epsilon | \bar{A})P(\bar{A}) \\ &= P(|Y - Z| \geq \epsilon) \\ &\rightarrow 0 \end{aligned}$$

Above all, we have $-\frac{2}{m}\sum_{i=1}^{k} logP_{(i)} \rightarrow \sum_{i=1}^{k} c_i$.

Now, consider the case of $K > k$ that K could cover k nonzero signals under the alternative and also include noises P_i, $i = k + 1, ..., n$ under the null. we rewrite these ordered p-values into two parts:

$$-\frac{2}{m}\sum_{i=1}^{K} logP_{(i)} = -\frac{2}{m}\sum_{i=1}^{k} logP_i - \frac{2}{m}\sum_{i=1}^{K-k} logU_{(i)}$$

where $U_{(i)}, i = 1, ..., K-k$ is ordered statistics from $n-k$ iid random variables of $Unif(0,1)$.

With high probability, we already have $-\frac{2}{m}\sum_{i=1}^{k} logP_i \to \sum_{i=1}^{k} c_i(\theta)$. Now, focus on the convergence of $-\frac{2}{m}\sum_{i=1}^{K-k} logU_{(i)}$. Since k is a fixed constant and $n \to \infty$, the total number of noise $n-k$ can be approximated by n for simplicity. Then, calculate the convergence of $-\frac{2}{m}\sum_{i=1}^{K-k} logU_{(i)}$. Here, we employ the expected value of $\prod_{i=1}^{K-k} U_{(i)}$: Note that any uniform order statistics $U_{(i)}$ is represented as the product of powers of independent uniformly distributed random variables,

$$U_{(i)} = W_i^{1/i} W_{i+1}^{1/i+1} ... W_n^{1/n}, i = 1, 2, ..., n.$$

where W_i is independent uniformly distributed on [0,1] random variables(Ahsanullah et al., 2013). Then, define a random variable $Y = -2\log(\prod_{i=1}^{K-k} U_{(i)})/m$, and by the representative of uniform random variable,

$$Y = -2\log(W_1 ... W_{K-k} W_{K-k+1}^{K-k/K-k+1} ... W_n^{K-k/n})/m.$$

Let $X_1 = \frac{-2\log W_1}{m}, ..., X_{K-k} = \frac{-2\log W_{K-k}}{m}, X_{K-k+1} = \frac{-2(K-k)\log W_{K-k+1}}{m(K-k+1)}, ..., X_n = \frac{-2(K-k)\log W_n}{mn}$ are independent random variables, we have another expression of $Y = X_1 + X_2 + ... + X_n$.

In order to get the convergence of Y, we firstly derive the expected value and variance of Y:

$$
\begin{aligned}
E(Y) &= E(X_1 + X_2 + ... + X_n) \\
&= E(\frac{-2\log W_1}{m} + ... + \frac{-2\log W_{K-k}}{m} + \frac{-2(K-k)\log W_{K-k+1}}{m(K-k+1)} + ... + \frac{-2(K-k)\log W_n}{mn}) \\
&= E(\frac{-2\log W_i}{m})E(K-k + \frac{K-k}{K-k+1} + ... + \frac{K-k}{n}) \\
&= 2E(\frac{K-k}{m} + \frac{1}{m}\frac{K-k}{K-k+1} + ... + \frac{1}{m}\frac{K-k}{n})
\end{aligned}
\tag{3.7}
$$

Except for the first term in the parentheses of the above formula, the lower bound and upper bound for the rest terms are:

$$
\begin{aligned}
(n(1-\tau_1))E(\frac{1}{m}\frac{K-k}{K-k+1}) &= \frac{n(1-\tau_1)}{m}\sum_{k'=k}^{n}\frac{k'-k}{k'-k+1}\binom{n}{k'}\tau_1^{k'}(1-\tau_1)^{n-k'} \\
&< \frac{n(1-\tau_1)}{m}\sum_{k'=k}^{n}\binom{n}{k'}\tau_1^{k'}(1-\tau_1)^{n-k'} < \frac{1-\tau_1}{c}
\end{aligned}
\tag{3.8}
$$

$$
(n(1-\tau_1))E(\frac{1}{m}\frac{K-k}{n}) = \frac{n(1-\tau_1)(n\tau_1 - k)}{mn} \approx \frac{\tau_1(1-\tau_1)}{c} \tag{3.9}
$$

Thus,

$$
E(Y) = \begin{cases} 2\frac{n\tau_1 - k}{m} \to 0, \text{ when } n = o(m) \\ 2\frac{n\tau_1 - k}{m} \to \frac{2\tau_1}{c} + C, \text{ when } m = cn, C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c}.) \end{cases}
$$

Also, we study the variance of Y:

$$
\begin{aligned}
Var(Y) &= V(X_1 + X_2 + ... + X_n) \\
&= Var(\frac{-2\log W_1}{m} + ... + \frac{-2\log W_{K-k}}{m} + \frac{-2(K-k)\log W_{K-k+1}}{m(K-k+1)} + ... + \frac{-2(K-k)\log W_n}{mn}) \\
&= Var(\frac{-2\log W_i}{m}(K-k)) + Var(\frac{-2\log W_i}{m}(\frac{K-k}{K-k+1})) + ... + Var(\frac{-2\log W_i}{m}(\frac{K-k}{n}))
\end{aligned}
\tag{3.10}
$$

Since random variable W_i and K are independent,

$$\begin{aligned}
&Var(\frac{-2\log W_i}{m}(K-k)) \\
&= V(\frac{-2\log W_i}{m})V(K-k)+E^2(\frac{-2\log W_i}{m})V(K-k)+V(\frac{-2\log W_i}{m})E^2(K-k) \\
&= \frac{4}{m^2}n\tau_1(1-\tau_1)+\frac{4}{m^2}n\tau_1(1-\tau_1)+\frac{4}{m^2}(n\tau_1-k)^2 \approx \frac{4}{m^2}(n\tau_1-k)^2
\end{aligned} \tag{3.11}$$

Similarly, for the second term in the formula of $Var(Y)$:

$$\begin{aligned}
&Var(\frac{-2\log W_i}{m}\frac{K-k}{K-k+1}) \\
&= V(\frac{-2\log W_i}{m})V(\frac{K-k}{K-k+1})+E^2(\frac{-2\log W_i}{m})V(\frac{K-k}{K-k+1})+V(\frac{-2\log W_i}{m})E^2(\frac{K-k}{K-k+1}) \\
&< \frac{4}{m^2}+\frac{4}{m^2}+\frac{4}{m^2} \approx 0
\end{aligned} \tag{3.12}$$

With general weak law of large number in (Resnick, 1998) on page 205,

Theorem 9 (General weak law of large numbers)**.** *Suppose* $\{X_n, n \geq 1\}$ *are independent random variables and define* $S_n = \sum_{j=1}^n X_j$*. If*

$$\sum_{j=1}^n P[|X_j| > n] \to 0$$

$$\frac{1}{n^2}\sum_{j=1}^n EX_j^2 1_{[|X_j| \leq n]} \to 0$$

then if we define

$$a_n = \sum_{j=1}^n E(X_j 1_{[|X_j \leq n|]})$$

we get

$$\frac{S_n - a_n}{n} \to 0.$$

Since X_i are independent random variables, $|X_i| < n, i = 1, ..., n,$ then

$\sum_{i=1}^n P(|X_i| > n) = 0.$

Also, $\frac{1}{n^2}\sum_{i=1}^n EX_i^2 I_{|X_j|\le n} = \frac{1}{n^2}\sum_{i=1}^n EX_i^2 = \frac{1}{n^2} = \frac{1}{n^2}\sum_{i=1}^n (V(X_i) + E^2(X_i))$,

In (3.8), (3.9), (3.11) and (3.12), we have $V(X_i) + E^2(X_i) \to 0$. Then, $\frac{1}{n^2}\sum_{i=1}^n EX_i^2 I_{|X_j|\le n} \to 0$.

By theorem 9,

$$Y = -\frac{2}{m}\sum_{i=1}^{K-k} logU_{(i)} \to E(Y).$$

Above all, under the alternative,

$$-\frac{2}{m}\sum_{i=1}^{K} logP_{(i)} \to \begin{cases} \sum_{i=1}^k c_i(\theta), \text{when } n = o(m) \\ \sum_{i=1}^k c_i(\theta) + \frac{2\tau_1}{c} + C, \text{when } m = cn \text{ and fixed } C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c}). \end{cases}$$

□

Based on lemma 4 and 5, we have:

Theorem 10. *Under the alternative(1.4), the convergence of T_S/m for TFisher (3.4) is*

$$\frac{T_S}{m} = -\frac{2}{m}\sum_{i=1}^{K}\log P_{(i)} + \frac{2K}{m}\log\tau_2 \to \begin{cases} \sum_{i=1}^k c_i(\theta), \textit{when } n = o(m) \\ \sum_{i=1}^k c_i(\theta) + \frac{2\tau_1}{c} + \frac{2\tau_1}{c}\log\tau_2 + C, \\ \textit{when } m = cn \textit{ and } C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c}). \end{cases} \tag{3.13}$$

Note that according to the definition of Bahadur exact slope that the slopes are nonnegative, τ_2 need to satisfy $\tau_2 \ge e^{-1+\sum_{i=1}^k \frac{\log P_i}{\tau_1 n}}$.

3.2.2 Bahadur Exact Slope for TFisher

We firstly introduce a lemma for deducing the exact slope of TFisher:

Lemma 6. *The binomial coefficient* $\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i}$ *gets the biggest value when* $i_0 = \frac{(n+1)\tau_2}{\tau_2-\tau_1+1}$.

Proof. The binomial coefficient gets the biggest value when the ratio of i^{nd} and $(i-1)^{nd}$ term equals to one. That is,

$$\frac{\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i}}{\binom{n}{i-1}\tau_2^{i-1}(1-\tau_1)^{n-i+1}} = \frac{\tau_2(n-i+1)}{(1-\tau_1)i} = 1,$$

$$i = \frac{(n+1)\tau_2}{\tau_2-\tau_1+1}.$$

□

Here, the lower and upper bounds of Bahadur exact slope for TFisher are given as follows:

Theorem 11. *When* $n = o(m)$, *the exact slope of TFisher is*

$$\sum_{i=1}^{k} c_i(\theta);$$

When $m = cn$, c *is a positive constant and* $n \to \infty$, *the lower and upper bounds for the exact slope of TFisher are* $2f_l(t)$ *and* $2f_u(t)$, *respectively, where*

$$t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau_1}{c} + \frac{2\tau_1}{c}\log\tau_2 + C \text{ and fixed } C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c}), \tag{3.14}$$

$$f_l(t) = \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct}{2} + \log\frac{\tau_1}{\tau_2}) + 1] - \frac{1}{c}\log(\tau_2-\tau_1+1), \tag{3.15}$$

and

$$f_u(t) = \begin{cases} \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct+\log(\tau_1/\tau_2)}{2}) + 1] - \frac{1}{c}\log(-\tau_1+1), & -\tau_1+1 < \tau_2 \\ \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct+\log(\tau_1/\tau_2)}{2}) + 1] - \frac{1}{c}\log\tau_2, & \textit{otherwise} \end{cases}. \tag{3.16}$$

Proof. Under the null hypothesis,

$$\begin{aligned}
&-\frac{1}{m}log(1-F(mt))\\
&=-\frac{1}{m}logP(T_S\geq mt) \text{ by } (3.5)\\
&=-\frac{1}{m}log((1-\tau_1)^n I_{\{mt\leq 0\}}+e^{-\frac{mt}{2}}\sum_{i=1}^{n}\sum_{j=0}^{i-1}\frac{(mt+2ilog(\tau_1/\tau_2))^j}{2^j j!}\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i})\\
&=\frac{1}{2}t-\frac{1}{m}log(\sum_{i=1}^{n}\sum_{j=0}^{i-1}\frac{(mt+2ilog(\tau_1/\tau_2))^j}{2^j j!}\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i})
\end{aligned}$$

Since $t_0=mt$ and t (3.13) is the nonnegative convergence value for $\frac{T_m}{m}$ under H_1, the indicator function equals to zero. Then,

$$-\frac{1}{m}log(1-F(mt))=\frac{1}{2}t-\frac{1}{m}\log\Big(\sum_{i=1}^{n}\frac{(mt+2ilog(\tau_1/\tau_2))^{i-1}}{2^{i-1}(i-1)!}\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i}\Big)$$

Because the binomial coefficient $\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i}$ gets the biggest value when $i_0=\frac{(n+1)\tau_2}{\tau_2-\tau_1+1}$ in Lemma 6, we have

$$\sum_{i=1}^{n}\frac{(mt+2ilog(\tau_1/\tau_2))^{i-1}}{2^{i-1}(i-1)!}\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i}\leq\sum_{i=1}^{n}(\frac{(mt+2ilog(\tau_1/\tau_2))^{i-1}}{2^{i-1}(i-1)!})\binom{n}{i_0}\tau_2^{i_0}(1-\tau_1)^{n-i_0}$$

The lower bound of f function is

$$-\frac{1}{m}log(1-F(mt))\geq\frac{1}{2}t-\frac{1}{m}\log\sum_{i=1}^{n}\Big(\frac{(mt+2ilog(\tau_1/\tau_2))^{i-1}}{2^{i-1}(i-1)!}\Big)-\frac{1}{m}\log\binom{n}{i_0}\tau_2^{i_0}(1-\tau_1)^{n-i_0} \tag{3.17}$$

Then, by lemma 2, the second term in the right hand side of equation (3.17) is

$$-\frac{1}{m}\log\sum_{i=1}^{n}\Big(\frac{(mt+2ilog(\tau_1/\tau_2))^{i-1}}{2^{i-1}(i-1)!}\Big)\to\begin{cases}-\frac{1}{c}(\log(ct/2+\log(\tau_1/\tau_2))+1), & n=cm\\ 0, & n=o(m)\end{cases}$$

And the last term in the right hand side of equation (3.17) is

$$-\frac{1}{m}\log\binom{n}{i_0}\tau_2^{i_0}(1-\tau_1)^{n-i_0} \to \begin{cases} -\frac{1}{c}\log(\tau_2-\tau_1+1), & n=cm \\ 0, & n=o(m) \end{cases}$$

The specific calculations for the above equation are shown below:

By Stirling's approximation,

$$\begin{aligned}
&-\frac{1}{m}\log\binom{n}{i_0} \\
&=-\frac{1}{c}[\log n-\frac{\tau_2}{\tau_2-\tau_1+1}\log\frac{(n+1)\tau_2}{\tau_2-\tau_1+1}-(1-\frac{\tau_2}{\tau_2-\tau_1+1})\log(n-\frac{(n+1)\tau_2}{\tau_2-\tau_1+1})] \\
&=-\frac{1}{c}[\frac{\tau_2}{\tau_2-\tau_1+1}\log\frac{\tau_2-\tau_1+1}{\tau_2}+(1-\frac{\tau_2}{\tau_2-\tau_1+1})\log\frac{\tau_2-\tau_1+1}{1-\tau_1}] \\
&=\frac{1}{c}\frac{\tau_2}{\tau_2-\tau_1+1}\log\frac{\tau_2}{\tau_2-\tau_1+1}+\frac{1}{c}(1-\frac{\tau_2}{\tau_2-\tau_1+1})\log(\frac{1-\tau_1}{\tau_2-\tau_1+1})
\end{aligned}$$

Also,

$$-\frac{1}{m}\log\tau_2^{i_0}(1-\tau_1)^{n-i_0}\to-\frac{1}{c}\frac{\tau_2}{\tau_2-\tau_1+1}\log\tau_2-\frac{1}{c}(1-\frac{\tau_2}{\tau_2-\tau_1+1})\log(1-\tau_1)$$

Thus, the convergence of the largest binomial term is:

$$\begin{aligned}
&-\frac{1}{m}\log\binom{n}{i_0}\tau_2^{i_0}(1-\tau_1)^{n-i_0} \\
&\to\frac{1}{c}\frac{\tau_2}{\tau_2-\tau_1+1}\log\frac{1}{\tau_2-\tau_1+1}+\frac{1}{c}(1-\frac{\tau_2}{\tau_2-\tau_1+1})\log\frac{1}{\tau_2-\tau_1+1} \\
&=\frac{1}{c}\log\frac{1}{\tau_2-\tau_1+1}.
\end{aligned}$$

Above all, the f function when $n=o(m)$ is

$$f(t)=\frac{1}{2}t$$

and

$$t = \sum_{i=1}^{k} c_i(\theta).$$

The $f_l(.)$ function when $m = cn$ is

$$f_l(t) = \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct + \log(\tau_1/\tau_2)}{2}) + 1] - \frac{1}{c}\log(\tau_2 - \tau_1 + 1)$$

and

$$t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau_1}{c} + \frac{2\tau_1}{c}\log \tau_2.$$

Similarly, the smallest binomial coefficient $\binom{n}{i}\tau_2^i(1-\tau_1)^{n-i}$ is $(1-\tau_1)^n$, when $-\tau_1+1 < \tau_2$; otherwise the smallest binomial coefficient is τ_2^n.

Thus, the convergence of the smallest binomial term is:

$$\begin{aligned} -\frac{1}{m}\log\binom{n}{i_0}\tau_2^{i_0}(1-\tau_1)^{n-i_0} &\to -\frac{1}{c}\log(1-\tau_1), \text{ when } -\tau_1+1 < \tau_2; \\ &\to -\frac{1}{c}\log \tau_2, \text{ otherwise.} \end{aligned}$$

Thus,

$$f_u(t) = \begin{cases} \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct+\log(\tau_1/\tau_2)}{2}) + 1] - \frac{1}{c}\log(-\tau_1 + 1), & -\tau_1 + 1 < \tau_2 \\ \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct+\log(\tau_1/\tau_2)}{2}) + 1] - \frac{1}{c}\log \tau_2, & \text{otherwise} \end{cases}.$$

$\square$

Compared with the Bahadur exact slopes for RTP and fisher's, the one for TFisher performs as same as these methods when $n = o(m)$. In this case, the exact slopes when $n, m \to \infty$ are considered in theorem 11.

Note that when $\tau_1 = \tau_2 = \tau$, the lower and upper bounds of Bahadur exact slope for soft-thresholding are $2f_l(t)$ and $2f_u(t)$:

$$t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau}{c} + \frac{2\tau}{c}\log\tau + C \text{ ,where fixed } C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c}), \tag{3.18}$$

$$f_l(t) = \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct}{2}) + 1], \tag{3.19}$$

and

$$f_u(t) = \frac{1}{2}t - \frac{1}{c}[\log(\frac{ct}{2}) + 1] - \frac{1}{c}\log\tau_2. \tag{3.20}$$

We also could get the lower and upper bounds of Bahadur exact slope for hard-thresholding

$$T_h = \sum_{i=1}^{n}(-2\log P_i)I(P_i \leq \tau),$$

when $\tau_1 = \tau$ and $\tau_2 = 1$,

$$f_u(t) = \frac{1}{2}t - \frac{1}{c}(\log\frac{ct + \log\tau}{2} + 1) - \frac{1}{c}\log(1-\tau),$$

$$f_l(t) = \frac{1}{2}t - \frac{1}{c}(\log\frac{ct + \log\tau}{2} + 1) - \frac{1}{c}\log(2-\tau).$$

The exact slope $2f_l(t) \leq c(\theta) \leq 2f_u(t)$, where $t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau}{c}$.

Similarly, we could get the exact Bahadur slope for Fisher's test statistics, since f function does not contain the term $-\frac{1}{m}\log\binom{n}{i_0}\tau_2^{i_0}(1-\tau_1)^{n-i_0}$, when $\tau_1 = \tau_2 = 1$. The slope for fisher's is:

$$2f(t) = t - \frac{2}{c}(\log\frac{ct}{2} + 1)$$

and

$$t = b(\theta) = \sum_{i=1}^{k} c_i(\theta) + \frac{2}{c} + C, \text{ where fixed } C \in (\frac{\tau_1(1-\tau_1)}{c}, \frac{1-\tau_1}{c}).$$

Chapter 4

Discussion and Future Studies

In this chapter, we summarize the Bahadur exact slopes for one signal and k signals and state the future study goals.

In figure 4.1, the fisher's method, max fisher's method and max inverse normal method perform equally for the case of one signal. The inverse normal method has a smaller slope than the others. Furthermore, when the number of tests n goes to infinity, the exact slope of inverse normal method goes to zero.

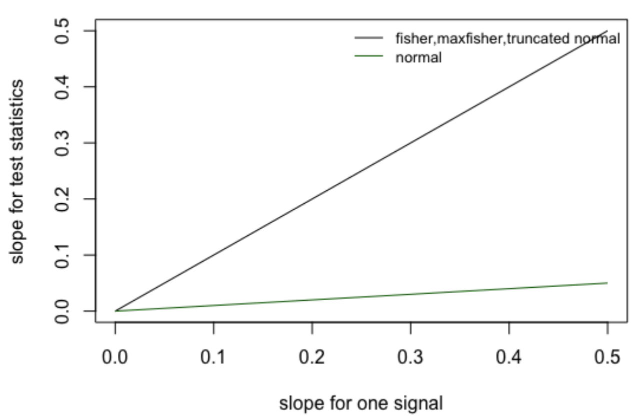

Figure 4.1: Exact slope for one signal. n=100.

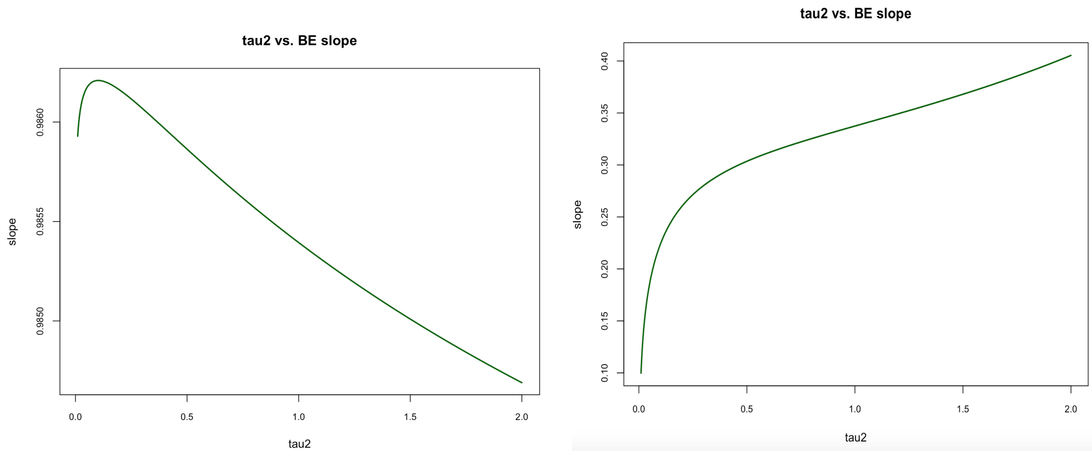

Figure 4.2: Bahadur exact slope over τ_2. Left panel: set $c = 1000$, $t = 1$ and $\tau_1 = 0.1$ and change τ_2 from 0.01 to 2. Right panel: set $c = 10$, $t = 0.6$ and $\tau_1 = 0.1$ and change τ_2 from 0.01 to 2.

Based on the lower bound of Bahadur exact slope for TFisher, we further study the choice of τ_1 and τ_2. There is no uniform rule for choosing τ_2 for different combinations of cs, ts. In figure 4, for the left one, the Bahadur exact slope attains the maximum when τ_2 is around τ_1 (i.e. soft-thresholding), while for the right one, the Bahadur exact slope attains the maximum when $\tau_2 = 1$ (i.e. hard-thresholding).

Moreover, the bigger the constant c is, the smaller the difference between soft-thresholding and hard-thresholding is, which is shown in the figure 4.3. Further, the difference between the lower bounds of soft-thresholding and hard-thresholding is smaller, when c gets bigger. When $c \to \infty$, there is no difference among fisher's method, soft-thresholding and hard-thresholding as $n = o(m)$, which is consistent with Littell's theory in 1973. Also, for p-combination methods such as fisher's, soft-thresholding and hard-thresholding, the bigger the c is, the higher the exact slope is. This could be easily understood from the perspective of signal, denser signals enjoy higher Bahadur exact slopes for fixed k signals. The conclusion could also be verified by the cases of one signal, where the exact slope of finite n is higher than the exact slope of infinite n.

The figure 4.3 shows the different cases with different cs: When $c = 1$, which means the sample size equals the number of tests, the soft-thresholding is superior to fisher's method when the slope of nonzero signals $\sum c_i(\theta) < 1.2$. Because the lower bound of soft-thresholding is higher than the fisher's and hard-thresholding. Also, when c=1, the exact slope of hard-thresholding is zero, i.e. $-\frac{2}{m}\log P_m = 0$, since the condition $t_0 + 2k\log(\frac{\tau_1}{\tau_2}) > 0$ does not be satisfied in (3.5) and the right tail probability is 1. When $c = 10$, the differences among fisher's, soft and hard-thresholding become smaller. If $0 < \sum c_i(\theta) < 0.61$, the soft-thresholding is the best among these three combination methods. If $0.61 < \sum c_i(\theta) < 0.72$, the hard-thresholding is the best among the three. Otherwise, soft-thresholding or fisher's method are worth considering. When $c = 20$, the exact slope of soft-thresholding could be superior to fisher's method when $\sum c_i(\theta) < 0.2$. The hard-thresholding is only better than the others when $0.3 < t < 0.32$, otherwise fisher's and soft-thresholding with equal τ_1, τ_2 would be better.

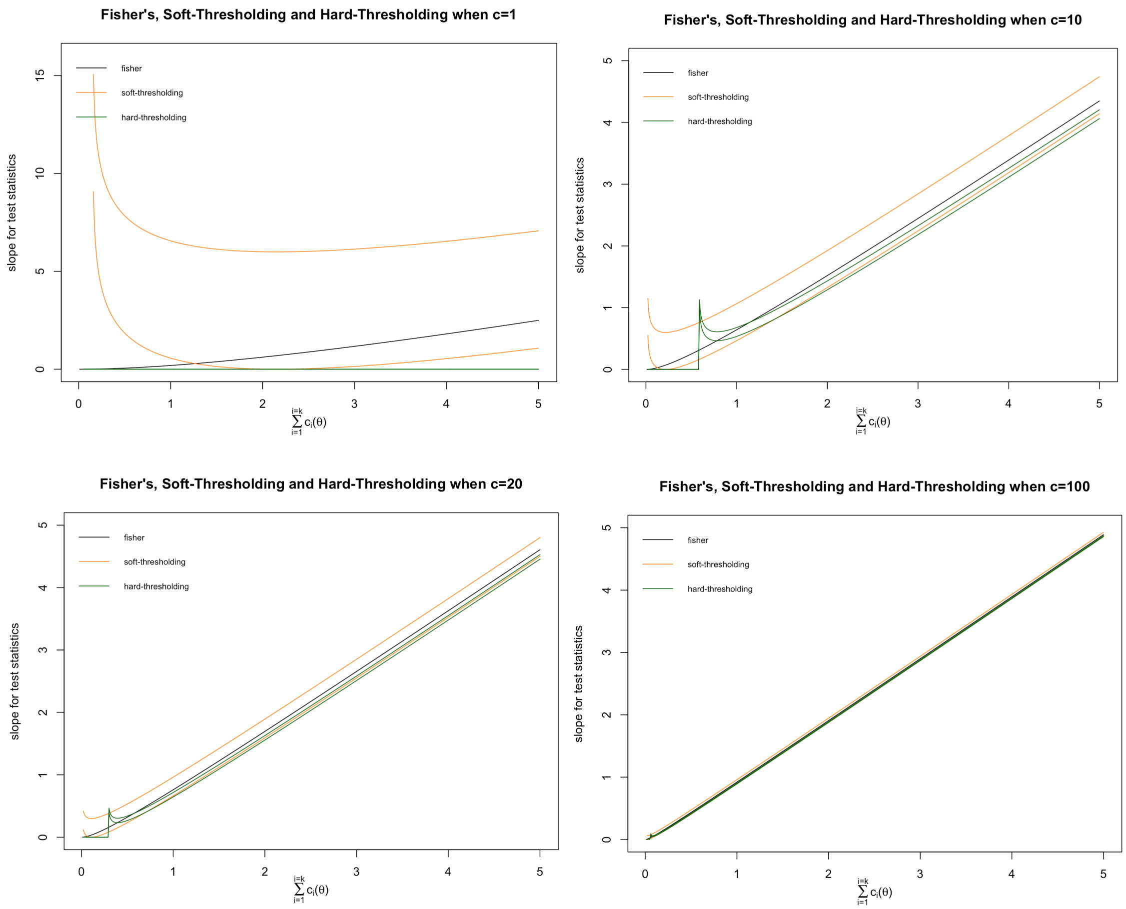

Figure 4.3: Orange line: the upper and lower bounds of the exact slope for soft-thresholding when $\tau_1 = \tau_2 = 0.05$. Black line: the exact slope for fisher's method when $\tau_1 = \tau_2 = 1$. Green line: the upper and lower bounds of the exact slope for hard-thresholding when $\tau_1 = 0.05$, $\tau_2 = 1$. C gets the value $\tau_1(1-\tau_1)/c$.

Here, plot the Bahadur exact slopes when $C = (1-\tau_1)/c$ in figure 4.4. From this figure, when $c = 1$ and $\sum c_i(\theta) < 0.5$, soft-thresholding is the best method, since the lower bound of soft-thresholding is higher than Fisher's and hard-thresholding. Also, when c increases, the Bahadur exact slope decreasing and the difference among Fisher's, Soft-Thresholding and Hard-Thresholding become small.

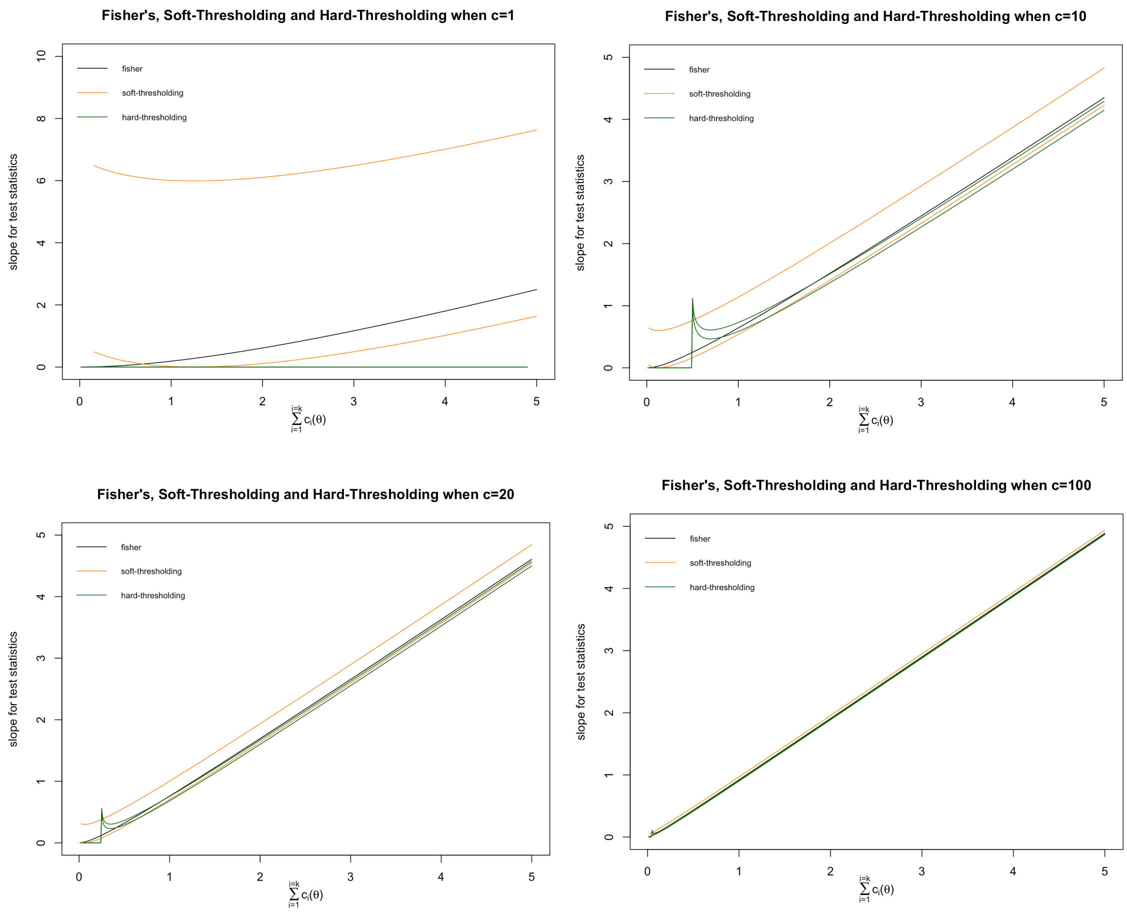

Figure 4.4: Orange line: the upper and lower bounds of the exact slope for soft-thresholding when $\tau_1 = \tau_2 = 0.05$. Black line: the exact slope for fisher's method when $\tau_1 = \tau_2 = 1$. Green line: the upper and lower bounds of the exact slope for hard-thresholding when $\tau_1 = 0.05$, $\tau_2 = 1$. C gets the value $(1 - \tau_1)/c$.

Throughout, the exact slopes for different case discussed before summarize in the following table 4:

The truncated inverse normal transformation method is superior to the non-truncated one for both finite and infinite number of hypothesis tests n, while it perform equally with the truncated log transformation method, for the case of one signal.

Test Statistics	H_1	Assumption	Exact Slope
$T_F = \sum\limits_{i=1}^{n} -2logP_{m_i}$	(1.3)	n is finite	$c_1(\theta)$
$T_{Fmax} = \max(-2\log P_{m_i})$	(1.3)	n is finite	$c_1(\theta)$
$T_F = \sum\limits_{i=1}^{n} -2logP_{m_i}$	(1.3)	$m = cn \to \infty$	$c_1(\theta) - \frac{2}{c}\Big[\log\big(c_1(\theta)c/2\big) + 1\Big]$
$T_F = \sum\limits_{i=1}^{n} -2logP_{m_i}$	(1.4)	$m = cn \to \infty$	$\sum_{i=1}^{k} c_i(\theta) + C - \frac{2}{c}\Big[\log\big(\sum_{i=1}^{k} c_1(\theta)(c+C)/2\big) + 1\Big]$
$T_{Fmax} = \max(-2\log P_{m_i})$	(1.3)	$n \to \infty$ and $n = o(m)$	$c_1(\theta)$
$T_N = \sum\limits_{i=1}^{n} Z_{m_i}$	(1.3)	n is finite	$\frac{c_1(\theta)}{n}$
$T_N = \sum\limits_{i=1}^{n} Z_{m_i}$	(1.3)	$n \to \infty$ and $n = o(m)$	0
$T_{Nmax} = \max Z_{m_i}$	(1.3)	for any $n \leq cm$	$c_1(\theta)$
$T_R = \sum_{i=1}^{k^*} -2logP_{(i)}$	(1.4)	n is finite or $n \to \infty$	$\sum_{i=1}^{k\wedge k^*} c_i(\theta)$, when $e^{-mt/2} \leq v^{k^*}$; 0, otherwise
$T_h = -2\log\prod_{i=1}^{n} P_i^{I(P_i\leq\tau)}$	(1.4)	$n \to \infty$ and $n = o(m)$	$\sum_{i=1}^{k} c_i(\theta)$
$T_h = -2\log\prod_{i=1}^{n} P_i^{I(P_i\leq\tau)}$	(1.4)	$m = cn \to \infty$	$2f_l(t) = t - \frac{2}{c}[\log(ct/2 + \log\tau) + 1] - \frac{2}{c}\log(2-\tau)$, where $t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau}{c} + C$.
$T_s = -2\log\prod_{i=1}^{n}\left(\frac{P_i}{\tau}\right)^{I(P_i\leq\tau)}$	(1.4)	$n \to \infty$ and $n = o(m)$	$\sum_{i=1}^{k} c_i(\theta)$
$T_s = -2\log\prod_{i=1}^{n}\left(\frac{P_i}{\tau}\right)^{I(P_i\leq\tau)}$	(1.4)	$m = cn \to \infty$	$2f_l(\theta) = t - \frac{2}{c}\Big[\log(ct/2) + 1\Big]$, where $t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau}{c} + \frac{2\tau}{c}\log\tau + C$.
$T_S = -2\log\prod_{i=1}^{n}\left(\frac{P_i}{\tau_2}\right)^{I(P_i\leq\tau_1)}$	(1.4)	$n \to \infty$ and $n = o(m)$	$\sum_{i=1}^{k} c_i(\theta)$
$T_S = -2\log\prod_{i=1}^{n}\left(\frac{P_i}{\tau_2}\right)^{I(P_i\leq\tau_1)}$	(1.4)	$m = cn \to \infty$	$2f_l(\theta) = t - \frac{2}{c}[\log(ct/2 + \log\frac{\tau_1}{\tau_2}) + 1] - \frac{2}{c}\log(\tau_2 - \tau_1 + 1)$, where $t = \sum_{i=1}^{k} c_i(\theta) + \frac{2\tau_1}{c} + \frac{2\tau_1}{c}\log\tau_2 + C$.

Table 4.1: Summary of the exact slopes. Note that for the case of hard-thresholding and soft-thresholding, the lower bounds of exact slope are used in the column Exact Slope.

For future studies, to be more accurate, we could further find the exact slopes for log-transformation methods, for example TFisher. Also, the truncated inverse normal transformation by threshold and rank should be studied from the perspective of Bahadur exact slope. The relationship between Bahadur efficiency and power could be studied further.

Appendix A

Appendix

A.1 Sub-exponential Distribution

Definition 1. *(Subexponential distribution function) Let X_i be iid postive rvs with df F such that $F(0) = 0$, $F(x) < 1$ for all $x > 0$, $F(\infty) = 1$. Denote*

$$\bar{F}(x) = 1 - F(x), x \geq 0$$

the tail of F and

$$\bar{F}^{n*}(x) = 1 - F^{n*}(x) = P(X_1 + X_2 + ... + X_n > x)$$

the tail of the n-fold convolution of F. F is a subexponential df ($F \in S$) if and only if one of the following equivalent conditions holds:

(a) $\lim_{x\to\infty} \frac{\bar{F}^{n*}(x)}{\bar{F}(x)} = n$ *for some(all)* $n \geq 2$,

(b) $lim_{x\to\infty} \int_0^\infty \frac{\bar{F}(x-t)}{\bar{F}(x)} dF(t) = 1$,

(c) $\lim_{x\to\infty} \frac{P(X_1+X_2+...+X_n>x)}{P(max(X_1,...,X_n)>x)} = 1$ *for some (all)* $n \geq 2$.

Conditions (a) and (c) were given by (Goldie and Klüppelberg, 1998); condition

(b) was given by (Pitman, 1980). The three conditions are equivalent.

Corollary 2. *Chi-square distribution is not included in the sub-exponential distribution.*

Proof. Let $X_1, X_2 \overset{ind}{\sim} \chi_2^2$ and $Y = X_1 + X_2 \sim \chi_4^2$, we have $\bar{F}_{X_1}(x) = e^{-\frac{x}{2}}$ and $\bar{F}_Y(y) = e^{-\frac{y}{2}} \sum_{i=0}^{1} \frac{y^i}{i!}$.

When $x = y \to \infty$,

$$\begin{aligned} lim_{x\to\infty} \int_0^x \frac{\bar{F}(x-t)}{\bar{F}(t)} dF(t) &= lim_{x\to\infty} \int_0^x \frac{e^{-(x-t)/2}}{e^{-x/2}} f(t)dt \\ &= lim_{x\to\infty} \int_0^x e^{t/2} \cdot \frac{1}{2} e^{-t/2} dt \\ &\neq 1 \end{aligned}$$

Besides, $\lim_{y\to\infty, x=y} \frac{e^{-\frac{y}{2}} \sum_{i=0}^{1} \frac{y^i}{i!}}{2e^{-\frac{x}{2}}} \neq 1$.

Thus, chi-square distribution is not included in the sub-exponential class. □

A.2 Some Deductions in Littell 1971

Here we give a clarification for the deduction of the fourth method in Littell 1971(Littell and Folks, 1971).

The overall test statistics is $T_n^{(m)} = -\frac{2}{\sqrt{n}} logminL_{n_i}^{(i)}$. Then

$$\begin{aligned} \frac{T_n^{(m)}}{\sqrt{n}} &= -\frac{2}{n} logminL_{n_i}^{(i)} \\ &= max(logL_{n_i}^{(i)}) \\ &\to max\lambda_i c_i(\theta) \end{aligned}$$

Under the null hypothesis, $-2logL_{n_i}^{(i)}$ follows a chi-square with 2 degress of freedom, which is exponential with parameter $\lambda = \frac{1}{2}$.

$$\begin{aligned}
-\frac{1}{n}log(1 - F_n^{(m)}(\sqrt{n}t)) &= -\frac{1}{n}log(1 - P(max(-\frac{2}{\sqrt{n}}logL_{n_i}^{(i)}) < \sqrt{n}t)) \\
&= -\frac{1}{n}log(1 - P(max(-2logL_{n_i}^{(i)}) < nt)) \\
&= -\frac{1}{n}log(1 - P(-2logL_{n_i}^{(i)}) < nt)^p) \\
&= -\frac{1}{n}log(1 - (1 - e^{-\frac{nt}{2}})^p) \\
&\to -\frac{1}{n}logpe^{-\frac{nt}{2}} \\
&\to \frac{t}{2}.
\end{aligned}$$

Bibliography

Walid A Abu-Dayyeh, Marwan A Al-Momani, and Hassen A Muttlak. Exact bahadur slope for combining independent tests for normal and logistic distributions. *Applied mathematics and computation*, 135(2):345–360, 2003.

Mohammad Ahsanullah, Valery B Nevzorov, and Mohammad Shakil. *An introduction to order statistics*. Springer, 2013.

Frank Dudbridge and Bobby PC Koeleman. Rank truncated product of p-values, with application to genomewide association scans. *Genetic epidemiology*, 25(4):360–366, 2003.

Ronald Aylmer Fisher. Statistical methods for research workers. 1934.

Charles M Goldie and Claudia Klüppelberg. Subexponential distributions. *A practical guide to heavy tails: statistical techniques and applications*, pages 435–459, 1998.

Nicholas A Heard and Patrick Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 2017.

James A Koziol and Henry C Tuckwell. A bayesian method for combining statistical tests. *Journal of statistical planning and inference*, 78(1):317–323, 1999.

Ramon C Littell and J Leroy Folks. Asymptotic optimality of fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66 (336), 1971.

Ramon C Littell and J Leroy Folks. Asymptotic optimality of fisher's method of combining independent tests ii. *Journal of the American Statistical Association*, 68(341):193–194, 1973.

Yakov Nikitin. *Asymptotic Efficiency of Nonparametric Tests*. Cambridge university press, 1995.

EJG Pitman. Subexponential distribution functions. *Journal of the Australian Mathematical Society (Series A)*, 29(30)(337-347), 1980.

Sidney Resnick. *A Probability Path*. 1998.

Michael C Whitlock. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *Journal of evolutionary biology*, 18(5): 1368–1373, 2005.

Hong Zhang, Tiejun Tong, John E Landers, and Zheyang Wu. Tfisher tests: Optimal and adaptive thresholding for combining p-values. *arXiv preprint arXiv:1801.04309*, 2018.