Calculating One-sided P-value for TFisher Under Correlated Data

by

Jiadong Fang

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

April 2018

APPROVED:

Professor Zheyang Wu, Major Thesis Advisor

Professor Luca Capogna, Head of Department

Abstract

P-values combination procedure for multiple statistical tests is a common data analysis method in many applications including bioinformatics. However, this procedure is non-trivial when input P-values are dependent. For the Fisher's combination procedure, a classic method is the Brown's Strategy [1, Brown,1975], which is based empirical moment-matching of gamma distribution. In this project, we address a more general family of weighting-and-truncation p-value combination procedures called TFisher. We first study how to extend Brown's Strategy to this problem. Then we make further development in two directions. First, instead of using the empirical polynomial model-fitting strategy to find moments, we developed an analytical calculation strategy based on asymptotic approximation. Second, instead of using the gamma distribution to approximate the null distribution of TFisher, we propose to use a mixed gamma distribution or a shifted-mixed gamma distribution. We focus on calculating the one-sided p-value for TFisher, especially the soft-thresholding version of TFisher. Simulations show that our methods much improve the accuracy than the traditional strategy.

Key Words: p-value combination test, one-sided p-value, TFisher, correlated data analysis

Acknowledgements

I would like to express my gratitude to my advisor Professor Wu with his patient and detailed instruction. I was greatly motivated to try my best and discovered useful conclusions through this project. He also gave me great support and encouragement when I met some problems, which made me more confident to solve them. With his instruction and encouragement, I finally complete the thesis.

My thanks also extend to Hong Zhang who helped me a lot on background knowledge learning. He kindly explained the theoretical background and also helped me on coding problems. Some of my thesis work is based on the former work of his paper.

Contents

List of Figures

Chapter 1

Introduction

In order to integrate the large and diverse datasets found in systems biology, it is common to combine P-values from multiple statistical tests. The earliest method to combine independent P-values is seen in the work of [Fisher, 1948]. [Brown, 1975] extended Fisher's Method to the case where P-values are assumed to be drawn from a multivariate normal distribution with a known covariance matrix. [Kost and McDermott, 2002] further extended Brown's Method analytically for unknown covariance matrices.

Combining a group of hypothesis tests $X_i, i = 1, ..., n$, we usually focus on the corresponding p-values $P_i, i = 1, ..., n$,to form a single statistic for testing the property of the whole group. For example, in the scenario of meta-analysis each test corresponds to one study, and a group of similar studies and their p-values are combined to exam the evidence of a common scientific hypothesis of these studies. In the scenario of signal detection, each test is for one factor. The p-values of a group of factors are combined to exam whether some of those factors are associated with a common response variable. Because each p-value provides information of one source (i.e., a study or a factor), p-value combination method can be considered as combining evidences to make a reliable conclusion for the whole group. The classic Fisher's combination test [1, Brown,1975] can be equivalently written as $T = \prod_{i=1}^{n} P_i \Leftrightarrow T' = -2\sum_{i=1}^{n} log(P_i)$. In Brown's result, it has been proved that

$W' \sim \chi^2_{2n}$. Fisher's test has many good properties. In terms the functional for transforming the p-values, log transformation is superior than other transformations [3, Hong,2017]. Now we extend to the soft-thresholding which is given by $T = \sum_{i=1}^{n} -2\log\left(\frac{P_i}{\tau}\right) I(P_i < \tau)$ and TFisher which is given by $T = \sum_{i=1}^{n} -2\log\left(\frac{P_i}{\tau_2}\right) I(P_i < \tau_1)$ with constraint $\tau_1 \leq \tau_2, \tau_2 \leq 1$ In order to study on optimality of these methods from theoretical perspective, the arguments for those choices varies from τ, τ_1, τ_2 we would like to propose a way to simulate the distribution of test statistic T.

Chapter 2

Polynomial Fitting

We aim to figure out the distribution or the approximate distribution of the following where soft-thresholding is added to the original data pattern

$$W = \sum_{i=1}^{n} -2\log\left(\frac{P_i}{\tau}\right) I(P_i < \tau)$$

where I is the indicator function. So the main concern is whether we can still apply Brown's result to find out $W^{'} \sim c\chi^2_{2n}$. In order to get an approximation of the distribution of $W^{'}$, we still use moments' method since it have been proved to have general good results for distribution approximation.

Follow the idea of Brown (1975), we can use a two-parameter scaled chi-square distribution (this is equivalent to use Gamma distribution) to approximate the null distribution of W. That is $P(W \leq w) \approx F_{\chi^2_d}(\frac{w}{c})$ where c and d are determined by matching the first two moments, i.e.

$$\begin{cases} \mu = cd \\ \sigma^2 = 2c^2 d \end{cases} \Longrightarrow \begin{cases} c = \frac{\sigma^2}{2\mu} \\ d = \frac{2(\mu)^2}{\sigma^2} \end{cases}$$

2.1 Variance estimation of soft thresholding with correlated input: polynomial fitting approach

Since we have already determined which type of distribution we would choose, the next problem is to find out the parameters of proposed distribution. Consider soft-thresholding with one-sided p-values coming from standard normal, $P_i = (1 - \Phi(Z_i))$, $i = 1, ..., n$,

$$W = \sum_{i=1}^{n} -2 \log \left(\frac{P_i}{\tau} \right) I(P_i < \tau).$$

We further assume Z_i's are multivariate normal and $cov(Z_i, Z_j) = \rho_{ij}$. The general idea of approximation is to match the moments of W with some known distribution.

2.1.1 Calculate the moments of W

For simplicity, write $Y_i = -2 \log \left(\frac{P_i}{\tau} \right) I(P_i < \tau)$. Regardless the correlation structure, $EY_i = 2\tau$, therefore, $EW = 2n\tau$. For the second moment $VarW = \sum_{i,j} cov(Y_i, Y_j)$, it is important to note that $cov(Y_i, Y_j)$ is a function of τ and ρ_{ij}, denote by $h(\tau, \rho)$.

One straightfoward idea is to find each $cov(Y_i, Y_j)$ by simulation, compare with numerical integration the simulation method saves plenty time as well as storage. Comparing with numerical method we need n^2 numerical integrations can be computationally burdensome and will face numerical problems when deal with extreme values. On the other side, we usually deal with the data which n is pretty large. The followings are the basic steps for finding out a reasonable polynomial to calculate the covariance.

1. For τ_i from 0.001 to 1 and ρ_j from -0.99 to 0.99, simulate (z_{1k}, z_{2k}), $k = 1, ..., 10^5$, from bivariate normal (Z_1, Z_2) with correlation ρ_j.

2. For each of these replicates, calculate the p-values and pair value of (y_{1k}, y_{2k}) after the log transformation, .

3. Find the sample covariance cov_{ij} between (y_{1k}) and (y_{2k}).

4. Build a polynomial regression model $lm(cov \sim poly(\tau + \rho))$.

Note that when $\rho_{ij} = 1$, i.e. $Y_i = Y_j$, then $cov(Y_i, Y_j) = Var(Y_i) = 4\tau(2 - \tau)$

We actually can find out some important properties for this certain polynomial: 1) Since our Y_is are non-negative, the fitting should have constraints such that the resulting estimates are non-negative. 2) Even if we have very large R^2, meaning the individual term's error might be small, but when calculating the variance estimate, we have to sum them all (n^2 terms) up, thus the final error could be large.

2.2 One side case of W

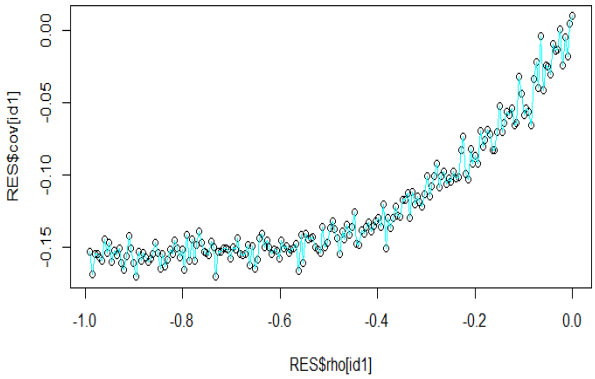
In order to randomly generate multi-dimensional Gaussian variables, we can either choose to use cholesky decomposition to get covariance matrix or MASS package to directly generate random variable with given correlation matrix.

The previous result for TFISHER is not satisfied in two parameters' cases, We try to figure out the relationship between covariance and τ_1, ρ given one of them fixed.

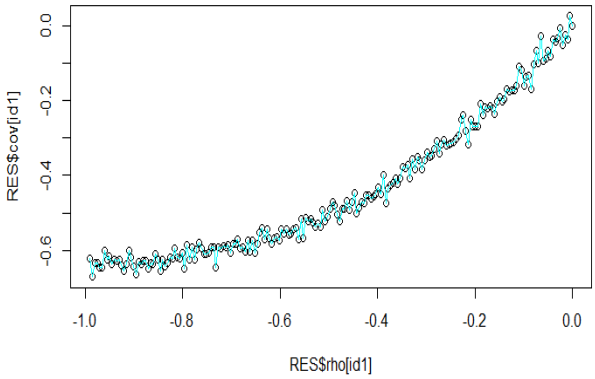
For the one side p-value case, the covariance pattern given τ_1 τ_2 fixed with respect to ρ is vary from a quadratic function to a strict line as Figure 2.1 shows when τ_1 increase from -0.99 to 0.

Figure 2.1: Polynomial Relationship between negative ρ and covariance

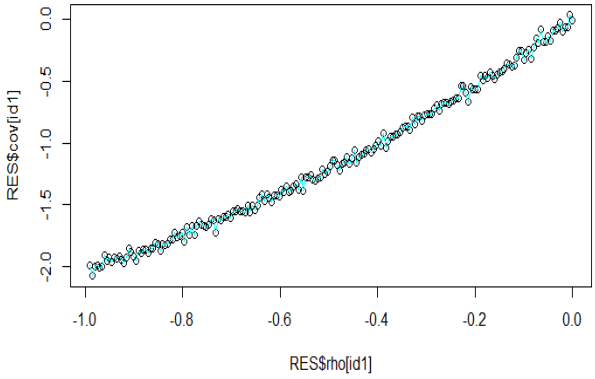
(a) $\tau_1 = \tau_2 = 0.01$



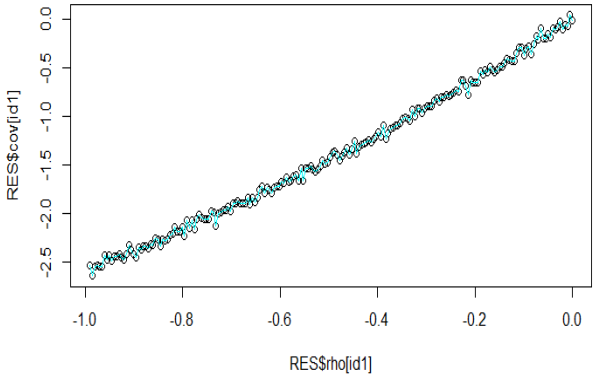
(b) $\tau_1 = \tau_2 = 0.1$



(c) $\tau_1 = \tau_2 = 0.5$



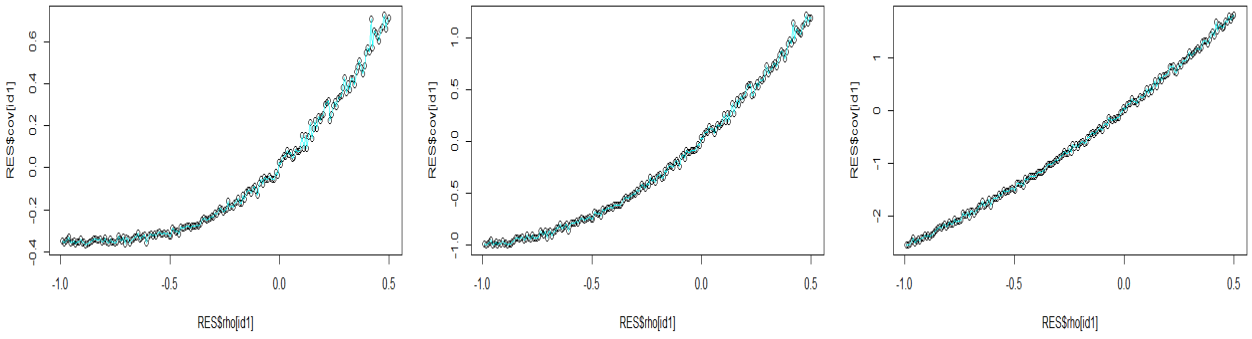
(d) $\tau_1 = \tau_2 = 0.9$



Also the intercept is increasing due to the truncation effect dies away since the formula has an identity function part which is explainable. We also try to combine the positive part of rho and given the range of rho from -0.99 to 0.5 from Figure2.2. The relationship remains the same.

Figure 2.2: Polynomial Relationship between wider range of ρ and covariance

(a) $\tau_1 = \tau_2 = 0.01$ (b) $\tau_1 = \tau_2 = 0.1$ (c) $\tau_1 = \tau_2 = 0.5$

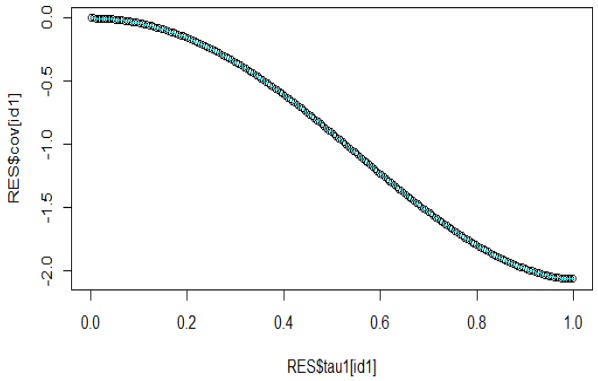


When we drop out the τ_1, τ_2 part since $\tau_1 = \tau_2 = 1$, then the relationship between covariance and rho has approximate line pattern as shown in Figure 2.2.

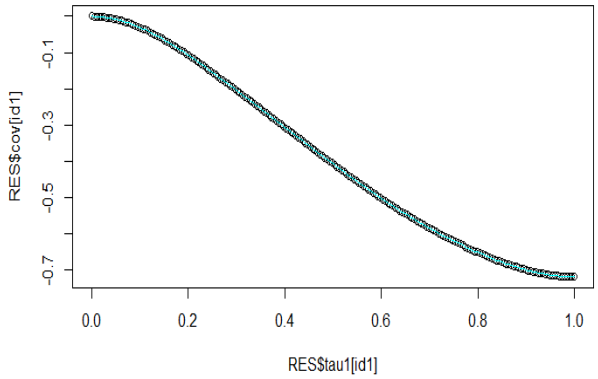
For the one side p-value case, the covariance pattern given ρ fixed with respect to $\tau_1\tau_2$ is vary from a quadratic function to a strict line when τ_1, τ_2 increase from 0.001 to 1. From Figure 2.3, it appears to be quite different patterns when ρ increases from negative to positive.

Figure 2.3: Polynomial Relationship between τ and covariance, here we assume $\tau_1 = \tau_2$

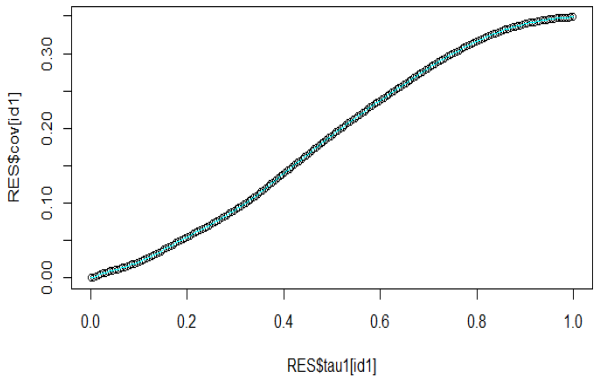
(a) $\rho = -0.9$



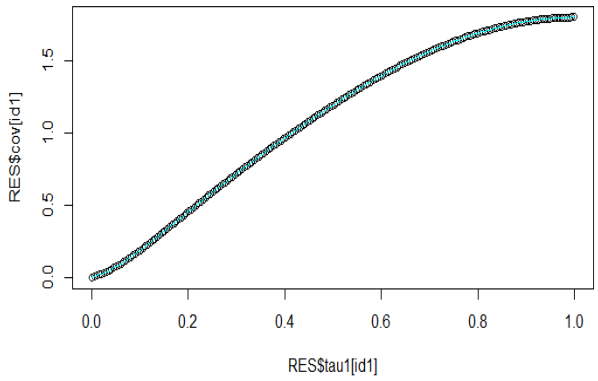
(b) $\rho = -0.5$



(c) $\rho = 0.5$



(d) $\rho = 0.9$



If adding all the possible parameters under 4^th order polynomial setting, $\mathbb{R}^2$ and $\mathbb{R}^2_a$ remain the same and most of the parameters still will be significant. This indicates adding more parameter terms cannot simply improve the fitting result.

We first try whether it can be shown as a symmetric case. However, the $\mathbb{R}^2$ will decrease 3% when we use the model fit with the negative ρ value and apply to the positive part. So we give formula for negative and positive part separately.

We give three candidate models which ρ would always be first order term. And the $\mathbb{R}^2$ for those two models have increased more than 1% from those model given ρ^2 terms.

For ρ from -0.99 to 0.99

1. $2.78\tau\rho + 4.1466\rho\tau^2 + 3.452\tau\rho^2 - 2.9187\rho^2\tau^2 - 3.648\rho\tau^3$ with $\mathbb{R}^2$ 97.59%.

2. $-0.187\tau_1 - 0.728\rho\tau_1 + 8.538\rho{\tau_1}^2 - 5.314\rho{\tau_1}^3$ with $\mathbb{R}^2$ 96.42%.

3. $(8\tau - 4\tau^2)\rho^2$ with $\mathbb{R}^2$ equals 92.51%.

Notice adding τ_1 term seems do not affect the parameters of $\rho{\tau_1}^2$ and $\rho{\tau_1}^3$ terms which simply change the parameter of $\rho\tau_1$.

Even we can simplify the model to case 3 which only contains two terms, it is interesting to notice the $\mathbb{R}^2$ does not decrease much which is actually more competitive. The model parameters change a lot is mainly caused by the negative sign of ρ. This could be a main problem when we choose models.

Actually, when ρ is positive, the quadratic model also can be applied. The $\mathbb{R}^2$ for those two models are very close. Even we can simplify the model to case 3 which only contains two terms, it is interesting to notice the $\mathbb{R}^2$ does not decrease much which is actually more competitive. For one side case, it seems to be robust within the range of ρ between -0.99 to 1. The certain pattern remains the same.

2.2.1 Comparison result of One-sided case

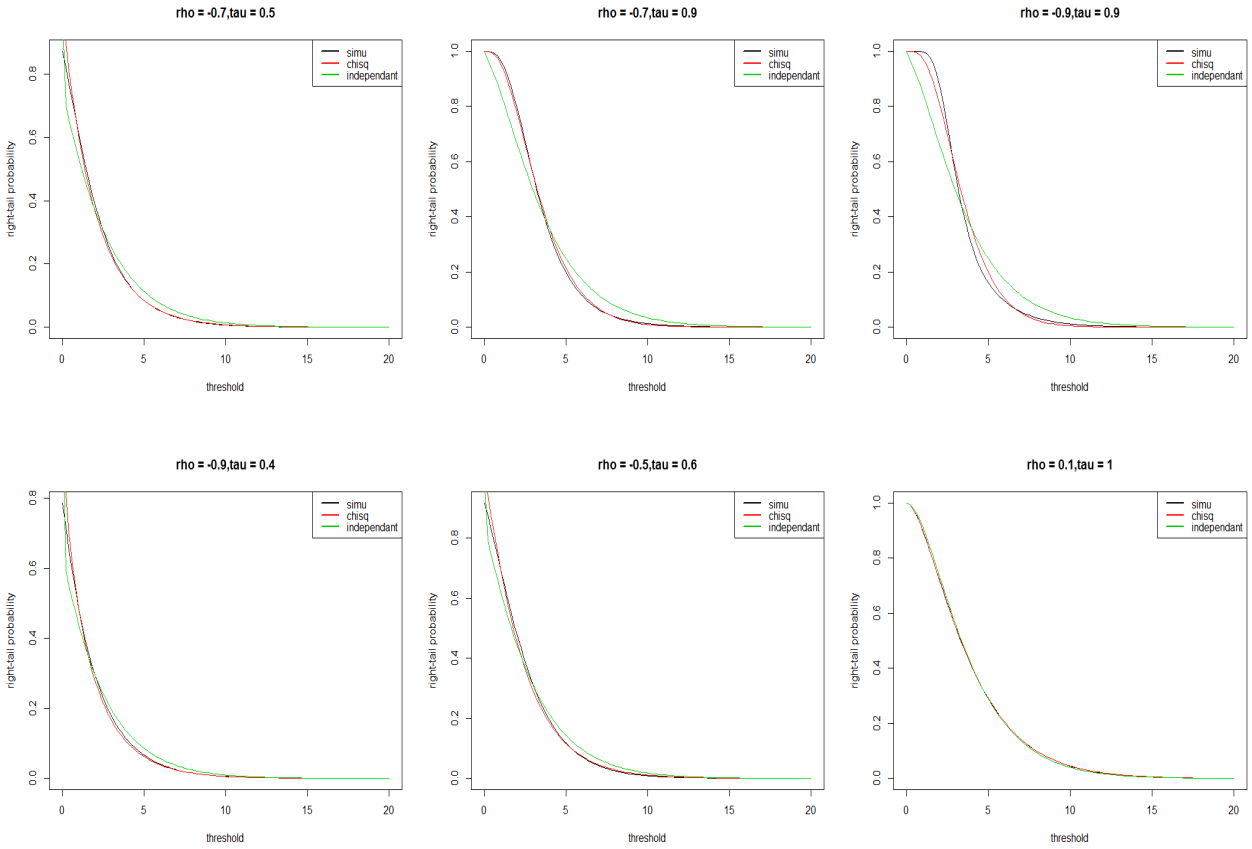
The other thing which we cannot neglect is that we need to test model stability for different truncation value τ_1 and small correlation ρ.

We have tried different candidate models from what we get. If we include τ^3 and only first term of ρ, the model becomes unstable in high negative correlated cases. The p-value of χ^2 will give error which due to the negative variance calculated.

Also the result of those including third order is not satisfying.

We try different models with both fourth order terms and high order intersection terms. When we including fourth order terms of τ_1 and ρ. The parameter of those high order terms is small comparing to those second order terms. Taking the range of τ_1 and ρ in to consideration, it should always belong to (0,1). We suggest the model below for the one-sided case. $2.78\tau\rho + 4.1466\rho\tau^2 + 3.452\tau\rho^2 - 2.9187\rho^2\tau^2 - 3.648\rho\tau^3$. Results are shown below:

Figure 2.4: P-Value graphs of different settings of ρ and τ of suggested model



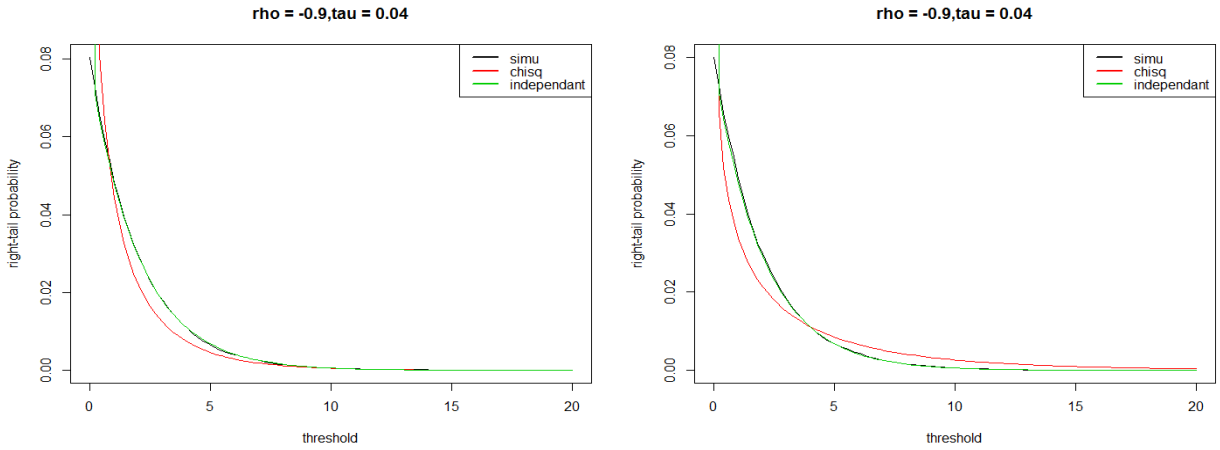
From Figure 2.4, the model works much better in high correlated cases with moderate truncation and small truncation of τ_1 compare to the independent case. However for small value of τ_1, the independent case works better. If we let $\tau = 1$ and small correlated cases,

the model also works well.

Figure 2.5: Comparison between two candidate models

(a) $(8\tau - 4\tau^2)\rho^2$

(b) $2.78\tau\rho + 4.1466\rho\tau^2 + 3.452\tau\rho^2 - 2.9187\rho^2\tau^2 - 3.648\rho\tau^3$



From Figure 2.5 when $\tau \leq 0.1$, $(8\tau - 4\tau^2)\rho^2$ model works better than suggested one. When for those $\tau \geq 0.2$ and $\rho \leq 0$, $2.78\tau\rho + 4.1466\rho\tau^2 + 3.452\tau\rho^2 - 2.9187\rho^2\tau^2 - 3.648\rho\tau^3$ works better. More important is in this case, the suggested model works much better than independent case.

2.2.2 Conclusion

1. Since the statistic W we consider is a function of τ_1, τ_2, ρ. So We consider the intercept term can be explained by a function of τ_1, τ_2,ρ. However in the model fitting, the intercept term will always be significant.

2. For one side case, what we get is different with Brown's paper [1, Brown,1975] since he included quadratic term to function of covariance and we explain the quadratic in τ_1,τ_2 part. Actually, we can notice for Brown's formula, the ρ^2 term has less effect

in independent cases and for the high correlated cases the curvature pattern does not significant in truncated cases.

3. For those τ_1 between (0.2,1) the truncated gamma works well for rho between (-0.99,-0.2) and (0.2,0.99). For those τ_1 between (0,0.2) the independent case works well.

4. It seems we cannot find a global optimal model for the two moments fit. Including high order term would significantly reduce the effectiveness for τ_1 between (0,0.2). On the other hand, the independent case has advantages when τ_1 and rho is small.

2.3 Comparison for gamma distribution and mixed-gamma distribution under polynomial fitting

Original two-parameter chi-square distribution is used to approximate the null distribution of W. The parameter estimates given below

$$P(W \leq w) \approx F_{\chi^2_d}(\frac{w}{c})$$

Using first and second memonts, we can find out the expression of c and d

$$\begin{cases} \mu = cd \\ \sigma^2 = 2c^2 d \end{cases} \Longrightarrow \begin{cases} c = \frac{\sigma^2}{2\mu} \\ d = \frac{2(\mu)^2}{\sigma^2} \end{cases}$$

Motivated by the truncation property of TFisher, we proposed a type of three-parameter mixed Gamma distribution: W is a random variable such that it has probability p_0 to be 0 and probability $1 - p_0$ to be a Gamma distribution. The Cumulative Distribution Function of this kind of distribution is $P(W \leq w) \approx p_0 + (1 - p_0)F_{\Gamma(k,\theta)}(w), w \geq 0$ [3, Hong,2017] where p_0 is the point mass probability at $w = 0$, k and θ are he shape and scale parameter

of Gamma distribution. p_0 can be estimated by $p_0 = P(P_i > \tau, i = 1, ..., n)$ which can be calculated by a multivariate normal distribution. k and θ are determined by matching the first two moments, i.e.

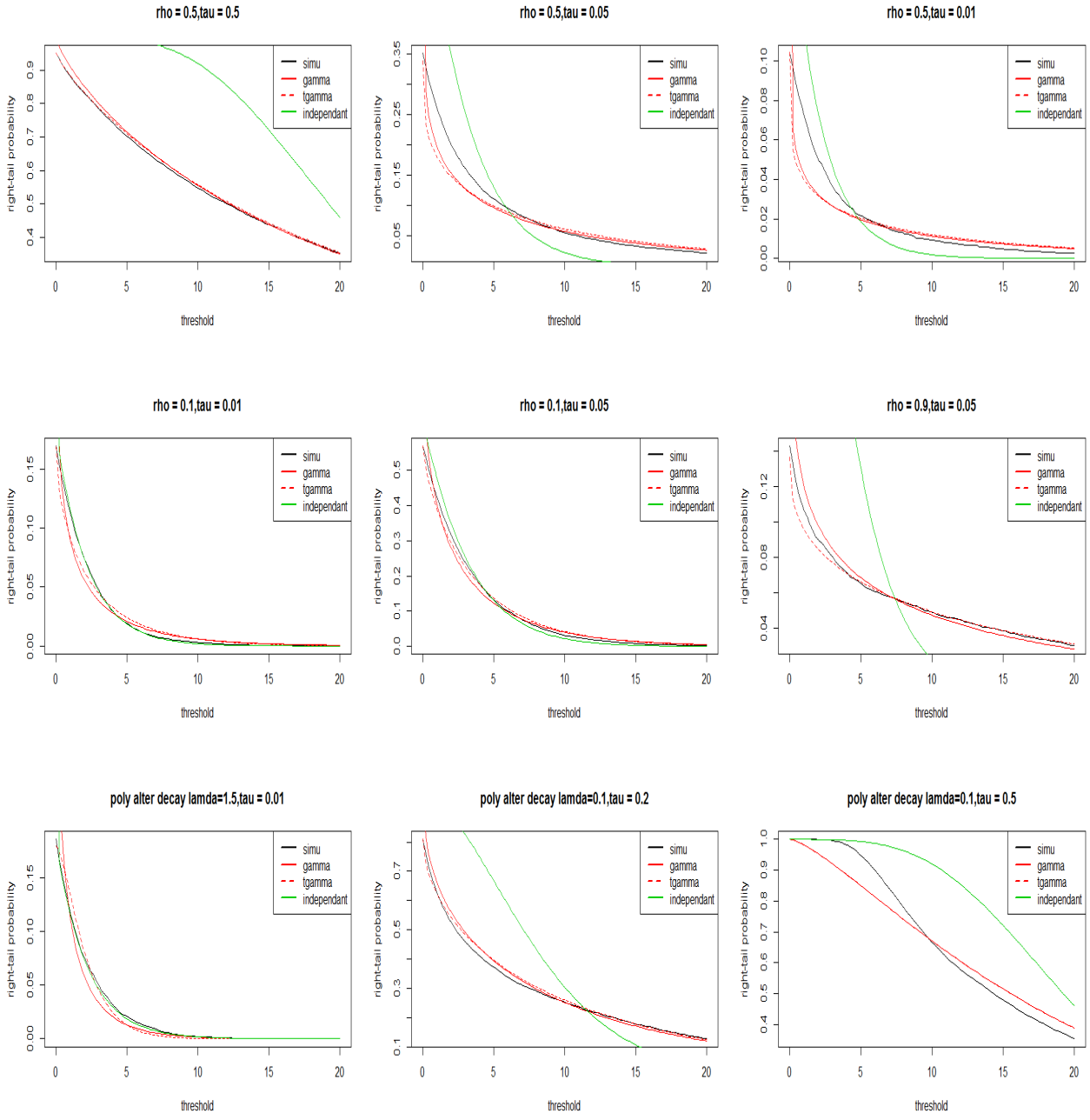
$$\begin{cases} \mu = (1 - p_0)k\theta \\ \sigma^2 = (1 - p_0)k\theta^2(kp_0 + 1) \end{cases} \Longrightarrow \begin{cases} k = \frac{\mu^2/\sigma^2}{1-p_0-p_0\mu^2/\sigma^2} \\ \theta = \frac{\sigma^2(1-p_0)-\mu^2 p_0}{(1-p_0)\mu} \end{cases}$$

Actually when the correlation is small, mixed Gamma approximation result is close to independence case, which is much better than we simply assume Gamma approximation. When the correlation is big, the proposed approximation is also decent.

In reality there are certain cases that both k and θ can be negative which would no longer be a Gamma distribution. One way to solve is to find out a nonnegative estimate for both k and θ. Or we may adjust p_0 instead of directly using the value let k, θ in the domain $(0, +\infty)$. In reality, that case does not appear throughout the whole simulation process.

Figure 2.6 are the graphs for comparison between original gamma distribution and truncated gamma distribution. The results follow what we discuss above.

Figure 2.6: Comparison between different distribution settings with suggested model



Chapter 3

Analytical Calculation for Soft Thresholding

Comparing with polynomial fitting method, we aim at finding a more general way to get correct variance which guarantees the correct moment estimation.

3.1 General Framework

Consider one-sided p-values coming from standard normal, $P_i = 1 - \Phi(Z_i)$, $i = 1, ..., n$, Z_i's are multivariate standard normal with correlation $\text{Cov}(Z_i, Z_j) = \rho_{ij}$. The soft-thresholding statistic is defined as

$$W = \sum_{i=1}^{n} -2 \log \left(\frac{P_i}{\tau} \right) I(P_i < \tau) = \sum_{i=1}^{n} Y_i$$

where $Y_i = -2 \log \left(\frac{P_i}{\tau} \right) I(P_i < \tau)$. The variance of W:

$$\text{Var}(W) = \sum_{i,j} \text{Cov}(Y_i, Y_j) = \sum_{i,j} \text{Cor}(Y_i, Y_j) \text{Var}(Y) = \text{Var}(Y) \left(\sum_{i,j} \text{Cor}(Y_i, Y_j) \right).$$

We already know exactly $\text{Var}(Y) = 4\tau(2 - \tau)$ thus no need to estimate it, so the only

thing left are those covariance terms, by substituting in: $\mathrm{Var}(W) = 4n\tau(2-\tau) + 4\tau(2-\tau)\sum_{i\neq j}\mathrm{Cor}(Y_i, Y_j)$

In order to calculating the variance part, we need to introduce several lemma with respect to asymptotic theory.

Lemma 1. *Given* $\tau \to 0$, $Y = -2\log\left(\frac{P}{\tau}\right) \approx Z^2 + \log Z^2 + \log 2\tau^2\pi$

Proof. By Mill's ratio $1 - \Phi(z) \overset{z\to\infty}{\to} \frac{\phi(z)}{z}$,

$$P = 1 - \Phi(Z) \approx \frac{1}{\sqrt{2\pi}}\frac{\exp(-Z^2/2)}{Z}$$

□

Lemma 1 shows that we may approximate Y by Z^2. Notice when Z goes to infinity, Z^2 gets dominated in above equation.

The next is a conjecture that supported by empirical evidences based on simulation results.

Conjecture 1. *Notice that* $P < \tau \iff Z > \Phi^{-1}(1-\tau)$, *write* $b = \Phi^{-1}(1-\tau)$, *we believe that*

$$\mathrm{Cor}(Y_i, Y_j) \approx \mathrm{Cor}\left(Z_i^2 I(Z_i > b), Z_j^2 I(Z_j > b)\right)$$

Then we can deduce analytical formula for the calculation of $\mathrm{Cor}(Y_i, Y_j)$. The formula will involve univariate integrals only which is feasible in reality even for large number of n.

3.1.1 Theoretical Deduction of $\mathrm{Cor}\left(Z_i^2 I(Z_i > b), Z_j^2 I(Z_j > b)\right)$

Since Z_i, Z_j are bivariate standard normal random variables with correlation ρ_{ij}, they have the same distribution as (U, V):

$$\begin{aligned} &U \sim N(0,1) \\ &V = \rho U + \sqrt{1-\rho^2} Z, \quad \rho = \rho_{ij} \end{aligned}$$

where Z is a standard normal random variable independent with U. Thus $\mathrm{Cor}\left(Z_i^2 I(Z_i > b), Z_j^2 I(Z_j > b)\right) = \mathrm{Cor}\left(U^2 I(U > b), V^2 I(V > b)\right)$. Under such transform, the underlying random variable U and Z are independent. [2, Casella,2002]

Next we will focus on the calculation of

$$\begin{aligned} &\mathrm{E}(U^2 I(U > b) V^2 I(V > b) \\ =&\mathrm{E}\left[U^2(\rho U + \sqrt{1-\rho^2} Z)^2 I(U > b) I(\rho U + \sqrt{1-\rho^2} Z > b)\right] \\ =&\mathrm{E}\left[\left(\rho^2 U^4 + 2\rho\sqrt{1-\rho^2} U^3 Z + (1-\rho^2) U^2 Z^2\right) I(U > b) I(\rho U + \sqrt{1-\rho^2} Z > b)\right]. \end{aligned} \tag{1}$$

The integration region is $\{b < u < +\infty\} \bigcup \{f(U) < z < +\infty\}$, where $f(u) = \frac{-\rho u + b}{\sqrt{1-\rho^2}}$. Denote $S = \left(\rho^2 U^4 + 2\rho\sqrt{1-\rho^2} U^3 Z + (1-\rho^2) U^2 Z^2\right)$.

Lemma 2. *Consider $Z \sim N(0,1)$. Let $M_n = \mathrm{E} Z^n I(c < Z < a)$. Then with $M_0 = \Phi(a) - \Phi(c)$, $M_1 = \phi(c) - \phi(a)$,*

$$M_n = (n-1) M_{n-2} - \left(a^{n-1}\phi(a) - c^{n-1}\phi(c)\right).$$

Proof.

$$\begin{aligned} M_n = \int_c^a z^n \phi(z) dz &= \int_c^a -z^{n-1} d\phi(z) \\ &= -z^{n-1} d\phi(z)|_a^c - (n-1) \int_c^a -z^{n-2} \phi(z) dz \\ &= (n-1) M_{n-2} - \Big(a^{n-1} \phi(a) - c^{n-1} \phi(c) \Big) \end{aligned}$$

□

We write the lemma 2 explicitly for the first four terms that we will infer later.

Corollary 1. *Let* $M_n(b) = \mathrm{E} Z^n I(b < Z < +\infty)$.

$$\begin{aligned} M_0(b) &= \Phi(+\infty) - \Phi(b) = \tau \\ M_1(b) &= \phi(b) \\ M_2(b) &= M_0 + b\phi(b) \\ M_3(b) &= 2\phi(b) + b^2 \phi(b) \\ M_4(b) &= 3M_2 + b^3 \phi(b) \end{aligned}$$

Corollary 2. *Let* $M_n(u) = \mathrm{E} Z^n I(f(u) < Z < +\infty)$. *Then*

$$\begin{aligned} &M_0(u) = \Phi(+\infty) - \Phi(f(u)) = 1 - \Phi(f(u)) \\ &M_1(u) = \phi(f(u)) \\ &M_2(u) = M_0(u) + f(u)\phi(f(u)) - f(+\infty)\phi(f(+\infty)) = M_0(u) + f(u)\phi(f(u)) \end{aligned}$$

By above moment deduction, we are able to get all the terms needed for covariance

calculation. Following are the terms needed in covariance calculation.

$$\begin{aligned}
&\mathrm{E}SI\left(b<U<+\infty\right)I(f(U)<Z<+\infty) \\
=&\int_b^{+\infty}\int_{f(u)}^{+\infty}\left(\rho^2u^4+2\rho\sqrt{1-\rho^2}u^3z+(1-\rho^2)u^2z^2\right)\phi(z)\phi(u)dzdu \\
=&\int_b^{+\infty}\left(\rho^2u^4M_0(u)+2\rho\sqrt{1-\rho^2}u^3M_1(u)+(1-\rho^2)u^2M_2(u)\right)\phi(u)du \\
=&\int_b^{+\infty}u^2h(u)du
\end{aligned} \tag{2}$$

where $h(u)\stackrel{def}{=}\left(\rho^2u^2M_0(u)+2\rho\sqrt{1-\rho^2}uM_1(u)+(1-\rho^2)M_2(u)\right)\phi(u)$. we conclude that

$$\begin{aligned}
&\mathrm{E}(U^2I(U>b)V^2I(V>b)) \\
=&\int_b^{+\infty}u^2h(u)du
\end{aligned} \tag{3}$$

$$\begin{aligned}
\mathrm{E}V^2I(V>b)&=\mathrm{E}U^2I(U>b) \\
&=M_2(b)=\tau+b\phi(b).
\end{aligned} \tag{4}$$

Therefore the covariance

$$\begin{aligned}
&\mathrm{Cov}(U^2I(U>b),V^2I(V>b)) \\
=&\int_b^{+\infty}u^2h(u)du-(\tau+b\phi(b))^2.
\end{aligned} \tag{5}$$

To find the correlation, we need to calculate the variance,

$$\begin{aligned}
\mathrm{Var}(U^2I(U>b))&=\mathrm{E}U^4I(U>b)-(\mathrm{E}U^2I(U>b))^2 \\
&=M_4(b)-(M_2(b))^2.
\end{aligned} \tag{6}$$

Finally, with $\mathrm{Var}Y = 4\tau(2-\tau)$, we have

$$\begin{aligned}
&\mathrm{Cov}(Y_i, Y_j) \\
=&\mathrm{Cor}(Y_i, Y_j)\mathrm{Var}(Y) \\
\approx&\mathrm{Cor}(U^2 I(U>b), V^2 I(V>b))\mathrm{Var}(Y) \\
=&\mathrm{Cov}(U^2 I(U>b), V^2 I(V>b))\frac{\mathrm{Var}(Y)}{\mathrm{Var}(U^2 I(U>b))} \\
=&\left[\int_b^{+\infty} u^2 h(u)du - (\tau + b\phi(b))^2\right]\frac{4\tau(2-\tau)}{M_4(b) - (M_2(b))^2}.
\end{aligned} \tag{7}$$

Equation (7) is the formula for variance calculation under soft-thresholding case.

3.1.2 Variance Calculation for under theoretical method

Following are the graphs for variance calculation under variety setting of τ and ρ. We fix τ value and let ρ varies from 0 to 1. For those ρ is less than 0, we use polynomial decaying covariance matrix.

From Figure 3.1 and 3.2, we can find out one-sided variance calculation is significantly proved compared with polynomial case when τ is relative small, i.e. in reality, most of time we would focus on small τ which is similar as significant level.

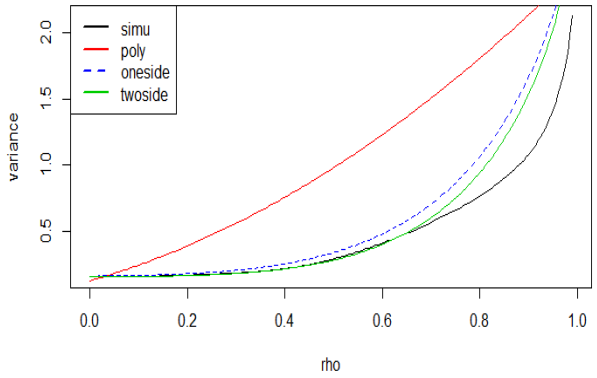
As τ gets larger, polynomial fitting gets better and closer to theoretical method. However, no overwhelming situation occurs since theoretical method perform consistently well under wide range of τ values. For the polynomial decaying situations, the results remain the same.

As the green lines shown in Figure 3.1 and 3.2, we can find direct extension to twoside case is not acceptable. We need to reconsider the integral domain[1].

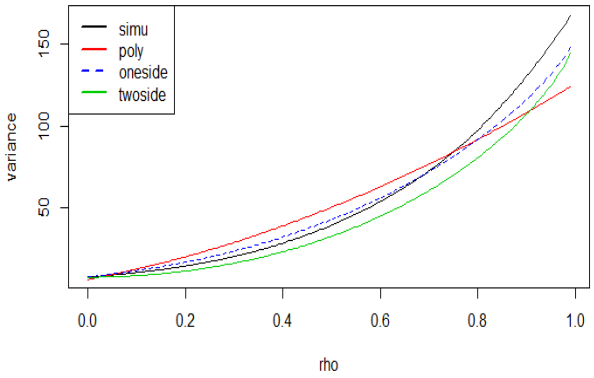
[1]This part has been down by Hong

Figure 3.1: Variance Calculation under equal correlation case

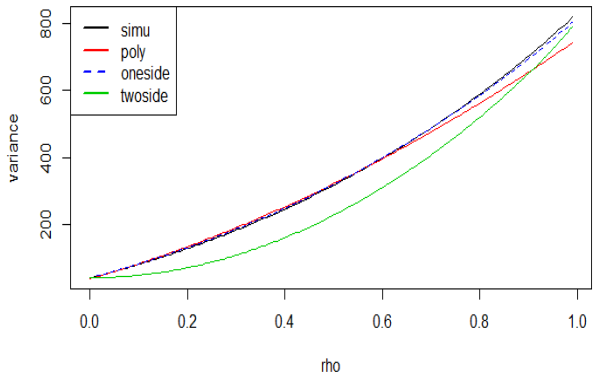
(a) $\tau = 0.001$



(b) $\tau = 0.05$



(c) $\tau = 0.3$



(d) $\tau = 0.5$

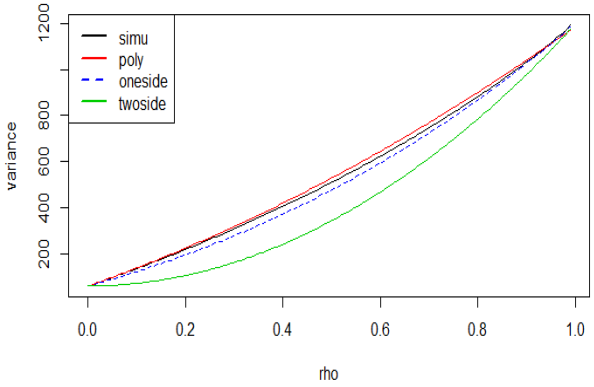
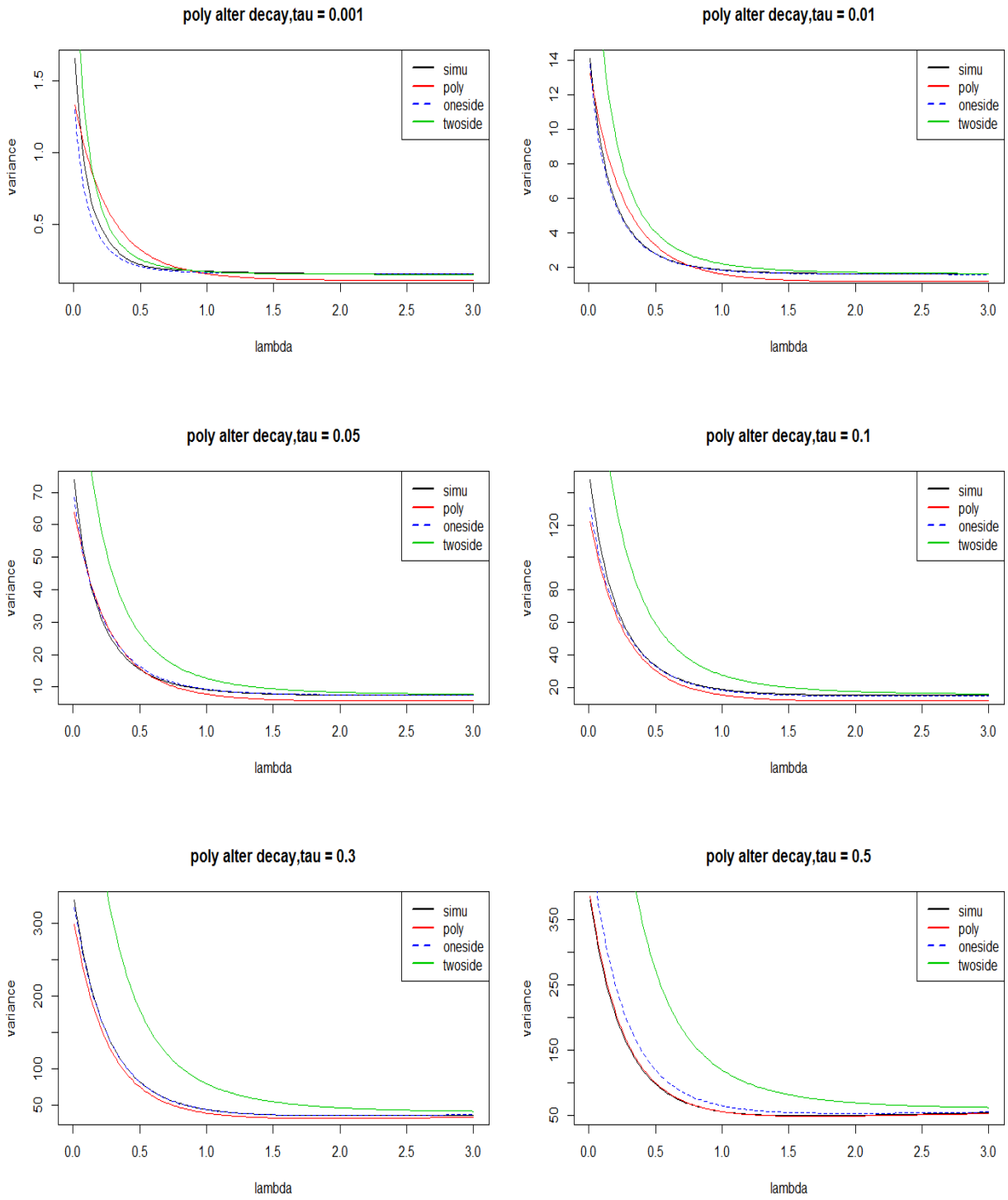


Figure 3.2: Variance under polynomial decaying correlation.



3.2 Extension formula for variance calculation of soft thresholding

The above calculation are under the assumption that we may approximate Y by Z^2. But in reality, the residual term will have influence for $Z \approx M$ where M is relatively small. Notice the second term $\log Z^2$We can find different function form or adding special term to compensate the residual effect. One direct way is that we choose $Z^2 + A$ instead of Z^2, where $A = F(\tau, E(\log Z^2 | Z > \Phi^{-1}(1-\tau)))$.

Lemma 3. *Given $\tau \to 0$, then $Y = -2\log\left(\frac{P}{\tau}\right) \approx Z^2 + \log Z^2 + \log 2\tau^2\pi$*

Proof. For general asymptotic expansion [5, Ober,2011], we have

$$P = 1 - \Phi(Z) \approx \frac{1}{\sqrt{2\pi}} \frac{\exp(-Z^2/2)}{Z} (1 - \frac{1}{Z^2} + \frac{3}{Z^4} + \cdots + \frac{(-1)^{n-1}(2n-3)!!}{Z^{2n-2}})$$

□

With the following conjecture:

Conjecture 2. *Since $P < \tau \iff Z > \Phi^{-1}(1-\tau)$, write $b = \Phi^{-1}(1-\tau)$, we believe that*

$$\mathrm{Cor}(Y_i, Y_j) \approx \mathrm{Cor}\left((Z_i^2 + A)I(Z_i > b), (Z_j^2 + A)I(Z_j > b)\right)$$

Similarly we can deduce analytical formula for the calculation of $\mathrm{Cor}(Y_i, Y_j)$.

3.2.1 Theoretical Deduction of $\mathrm{Cor}\left((Z_i^2+A)I(Z_i>b),(Z_j^2+A)I(Z_j>b)\right)$

Since Z_i, Z_j are bivariate standard normal random variables with correlation ρ_{ij}, and A is only a function of τ they have the same distribution as (U,V):

$$\begin{aligned} &U \sim N(0,1) \\ &V = \rho U + \sqrt{1-\rho^2}Z, \quad \rho = \rho_{ij} \end{aligned}$$

where Z is a standard normal random variable independent with U. Thus
$\mathrm{Cor}\left((Z_i^2+A)I(Z_i>b),(Z_j^2+A)I(Z_j>b)\right) = \mathrm{Cor}\left((U^2+A)I(U>b),(V^2+A)I(V>b)\right)$.
Under such transform, the underlying r.v. U and Z are independent.

Next we will focus on the calculation of

$$\begin{aligned} &\mathrm{E}[(U^2+A)I(U>b)(V^2+A)I(V>b)] \\ =&\mathrm{E}\left[(U^2+A)((\rho U+\sqrt{1-\rho^2}Z)^2+A)I(U>b)I(\rho U+\sqrt{1-\rho^2}Z>b)\right] \\ =&\mathrm{E}\left[\left(\rho^2U^4+2\rho\sqrt{1-\rho^2}U^3Z+(1-\rho^2)U^2Z^2+A(\rho U+\sqrt{1-\rho^2}Z)^2+AU^2+A^2\right)\right. \\ &\left. I(U>b)I(\rho U+\sqrt{1-\rho^2}Z>b)\right]. \end{aligned} \tag{8}$$

The integration region is $\{b<U<+\infty\}\bigcup\{f(U)<z<+\infty\}$, where $f(U)=\frac{-\rho U+b}{\sqrt{1-\rho^2}}$.

Denote $S=\left(\rho^2U^4+2\rho\sqrt{1-\rho^2}U^3Z+(1-\rho^2)U^2Z^2\right)$.

Corollary 1 and Corollary 2 are used in deduction.

Then the covariance calculation is different from previous cases, we write out each term

explicitly:

$$
\begin{aligned}
&\mathrm{E}SI\left(b<U<+\infty\right)I(f(U)<Z<+\infty) \\
=&\int_b^{+\infty}\int_{f(u)}^{+\infty}\left(\rho^2u^4+2\rho\sqrt{1-\rho^2}u^3z+(1-\rho^2)u^2z^2\right)\phi(z)\phi(u)dzdu \\
=&\int_b^{+\infty}\left(\rho^2u^4M_0(u)+2\rho\sqrt{1-\rho^2}u^3M_1(u)+(1-\rho^2)u^2M_2(u)\right)\phi(u)du \\
=&\int_b^{+\infty}u^2h(u)du
\end{aligned}
\tag{9}
$$

where $h(u)\overset{def}{=}\left(\rho^2u^2M_0(u)+2\rho\sqrt{1-\rho^2}uM_1(u)+(1-\rho^2)M_2(u)\right)\phi(u)$.

$$
\begin{aligned}
&\mathrm{E}[\frac{AS}{U^2}I\left(b<U<+\infty\right)I(f(U)<Z<+\infty)] \\
=&\int_b^{+\infty}\int_{f(u)}^{+\infty}\left(\rho^2u^2+2\rho\sqrt{1-\rho^2}uz+(1-\rho^2)z^2\right)A\phi(z)\phi(u)dzdu \\
=&\int_b^{+\infty}\left(\rho^2u^2M_0(u)+2\rho\sqrt{1-\rho^2}uM_1(u)+(1-\rho^2)M_2(u)\right)A\phi(u)du \\
=&\int_b^{+\infty}Ah(u)du.
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&\mathrm{E}AU^2I\left(b<U<+\infty\right)I(f(U)<Z<+\infty) \\
=&\int_b^{+\infty}\int_{f(u)}^{+\infty}AU^2\phi(z)\phi(u)dzdu \\
=&\int_b^{+\infty}AU^2M_0(u)\phi(u)du.
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
&\mathrm{E}A^2I\left(b<U<+\infty\right)I(f(U)<Z<+\infty) \\
=&\int_b^{+\infty}\int_{f(u)}^{+\infty}A^2\phi(z)\phi(u)dzdu \\
=&\int_b^{+\infty}A^2M_0(u)\phi(u)du.
\end{aligned}
\tag{12}
$$

we conclude that

$$\begin{aligned} &\mathrm{E}[(U^2+A)I(U>b)(V^2+A)I(V>b)] \\ =&\int_b^{+\infty}[u^2h(u)+Ah(u)+AU^2M_0(u)+A^2M_0(u)]du \end{aligned} \tag{13}$$

$$\begin{aligned} \mathrm{E}(V^2+A)I(V>b) &= \mathrm{E}(U^2I+A)I(U>b) \\ &= M_2(b)+AM_0(b) = \tau + b\phi(b) + A\tau. \end{aligned} \tag{14}$$

Therefore the covariance

$$\begin{aligned} &\mathrm{Cov}((U^2+A)I(U>b),(V^2+A)I(V>b)) \\ =&\int_b^{+\infty}[u^2h(u)+Ah(u)+AU^2M_0(u)+A^2M_0(u)]du-(\tau+b\phi(b)+A\tau)^2. \end{aligned} \tag{15}$$

To find the correlation, we need to calculate the variance,

$$\begin{aligned} \mathrm{Var}((U^2+A)I(U>b)) &= \mathrm{E}[U^4+2AU^2+A^2]I(U>b)-(\mathrm{E}(U^2I+A)I(U>b))^2 \\ &= M_4(b)+2AM_2(b)+A^2M_0(b)-(M_2(b)+AM_0(b))^2. \end{aligned} \tag{16}$$

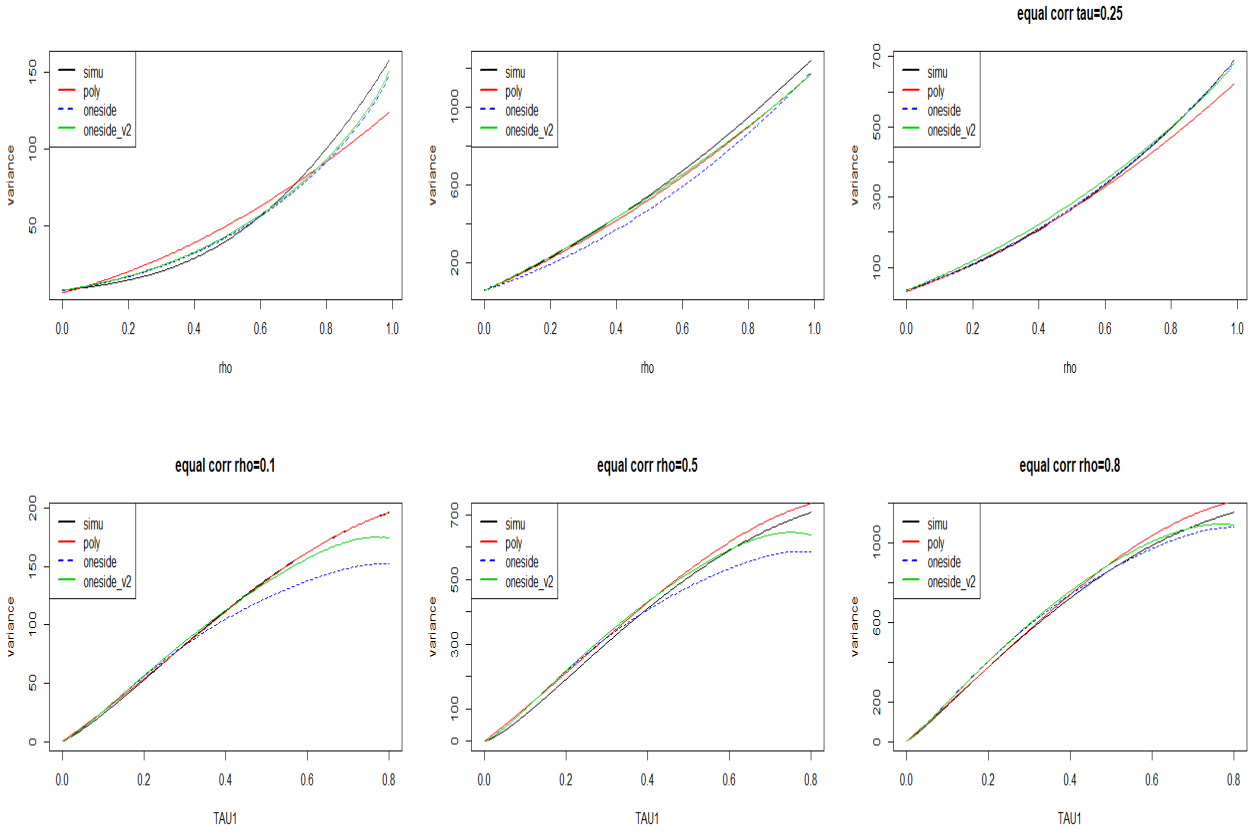
Finally, with $\mathrm{Var}Y = 4\tau(2-\tau)^2$ [4, Kost,2002], we have

$$\begin{aligned} &\mathrm{Cov}(Y_i,Y_j) \\ =&\mathrm{Cor}(Y_i,Y_j)\mathrm{Var}(Y) \\ \approx&\mathrm{Cor}((U^2+A)I(U>b),(V^2+A)I(V>b))\mathrm{Var}(Y) \\ =&\mathrm{Cov}((U^2+A)I(U>b),(V^2+A)I(V>b))\frac{\mathrm{Var}(Y)}{\mathrm{Var}((U^2I+A)I(U>b))} \\ =&\left[\int_b^{+\infty}[u^2h(u)+Ah(u)+AU^2M_0(u)+A^2M_0(u)]du-(\tau+b\phi(b)+A\tau)^2\right] \\ &\frac{4\tau(2-\tau)}{M_4(b)+2AM_2(b)+A^2M_0(b)-(M_2(b)+AM_0(b))^2}. \end{aligned} \tag{17}$$

[2]Theoretical value

3.2.2 Comparison for original method and extended method

Figure 3.3: Variance calculation with correction term by changing ρ



Here we give the variance calculation comparison graphs. The green line is denoted as oneside-v2 is under the extended method.

From Figure 3.3, we can find adding additional terms would improve the variance approximation especially when the correlation is large. The result are consistent with the graphs in section 3.1.2 when ρ increases the theoretical calculation of variance underestimate the true variance which is the simulation line.

By adding an additional term in theoretical calculation can be more flexible and adjust

the variance calculation between only containing Z^2 term and oracle ones. The extended form improves the accuracy of variance estimation as well as p-value calculation. Though finding a reasonable term A would be still a further research topic.

3.3 P-value calculation of soft-thresholding statistic

The following graphs 3.4 and 3.5 show the distribution of Fisher statistic W. We assume the empirical distribution which is the simulation one would always show reality case. The tgamma-oracle stands for those we already know the true variance and fit the truncated gamma distribution. The tgamma-poly stands for the polynomial fitting under the truncated gamma distribution. The tgamma-onesided stands for the extended theoretical method under the truncated gamma distribution. The independent line is based on the model without correlation.

We can find out the one-sided line is close to the oracle line (red one) in most cases. This result is consistent with the variance calculation results. Actually the truncated gamma distribution can be only determined by the first two moments which in this case is the mean and variance. Since the variance calculation is pretty well in theoretical methods so as the p-value calculation.

The polynomial fitting has large departure from the oracle one when τ is small which means we have small truncated value. In that case, polynomial is no longer accurate. This disadvantage disappears when τ increases and finally it gets closer to the simulation one. This drawback is mainly caused by fitting strategy. Since the model only consider polynomial terms and fit a wide range of τ and ρ. Actually when τ is small the variance is very small compare to those τs close to 1. So actually for small τ, polynomial fitting cannot tell what model it is.

However, we could get close fitting to oracle line. One thing we cannot neglect is oracle line itself still has departure from the simulation line, especially in the cases when τ is

small. We cast doubt on the truncated gamma distribution for small τ and probably can be improved by considering high order moments.

Figure 3.4: P-value calculation of soft-thresholding statistic under equal correlation. The ρ parameter is chosen such that the theoretical calculation has large deviation from the sample variance.

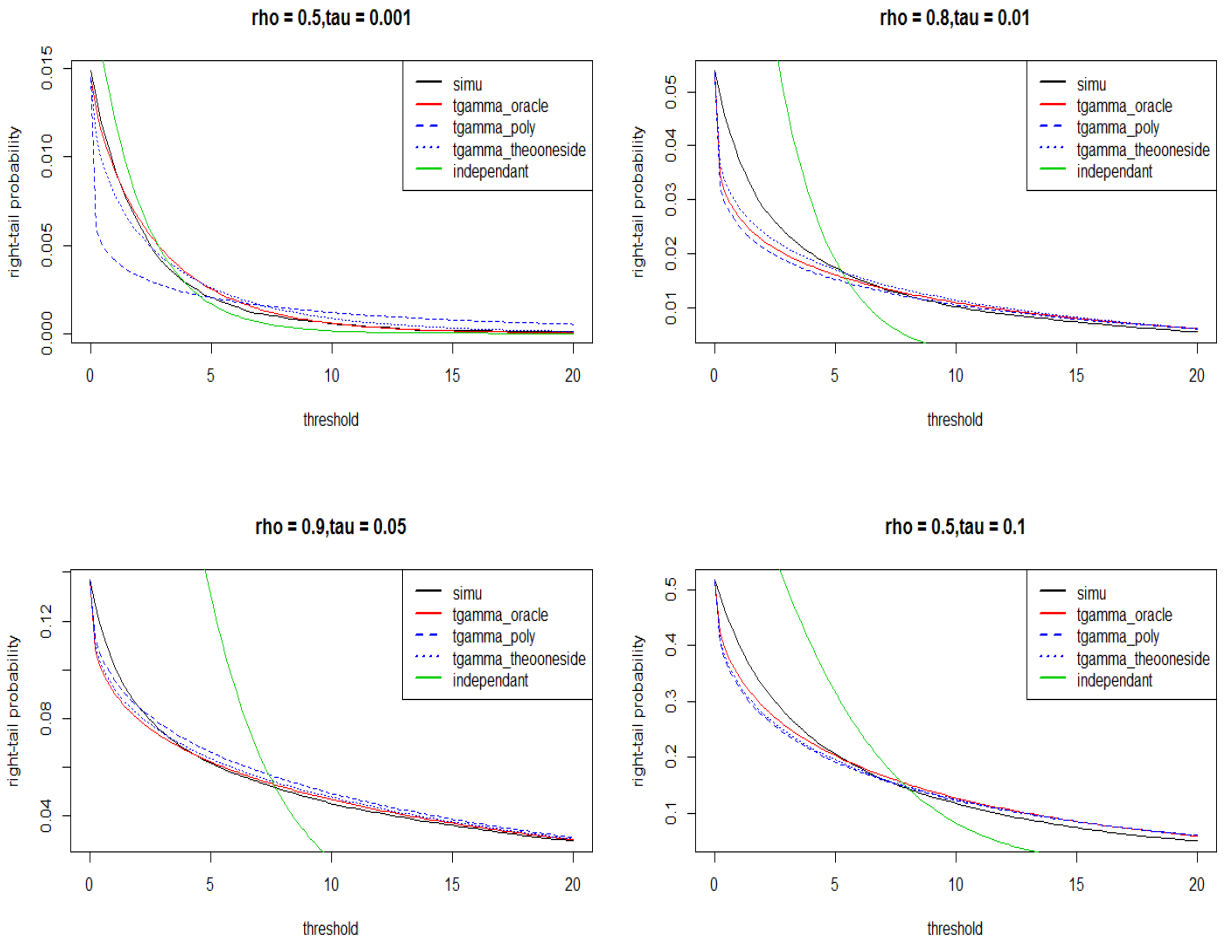
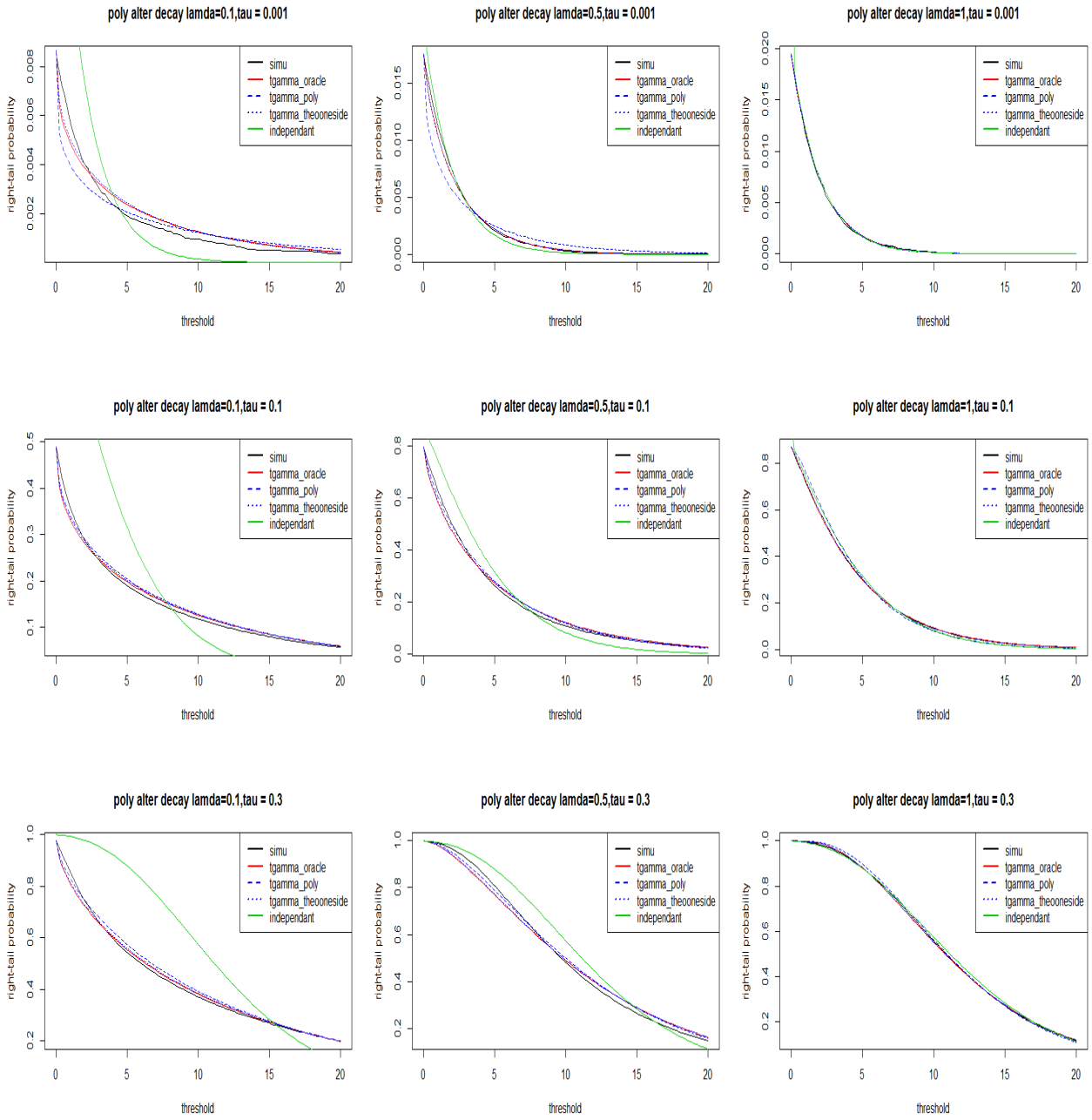


Figure 3.5: P-value calculation of soft-thresholding statistic under polynomial decaying correlation.



Chapter 4

Analytical Calculation for omnibus TFisher

4.1 Correlated TFisher: P-value and Omnibus Test

4.1.1 Shifted-Mixed Gamma

Since TFisher test statistic is based on truncated sum[1]. The nature way to deal with this problem is introduce a similar mixed gamma distribution but adding truncation property. One candidate model is (four-parameter) **shifted-mixed Gamma distribution** [7, Zaykin,2002]: W is a random variable such that it has probability p_0 to be 0 and probability $1-p_0$ to be a s-shifted Gamma random variable. The CDF of this kind of distribution is

$$P(W \leq w) \approx p_0 + (1-p_0)F_{\Gamma(k,\theta)}(w-s), w \geq 0$$

where p_0 is the point mass probability at $w=0$, k and θ are the shape and scale parameter of Gamma distribution.

Shift parameter s is the gap of discontinuity of a general TFisher statistic. It can be

[1]Here the statistic is actually discontinuous

estimated by

$$s = -2\log(frac\tau_1\tau_2)$$

p_0 can be estimated by

$$p_0 = P(P_i > \tau, i = 1, ..., n)$$

which can be calculated by a multivariate normal distribution. k and θ are determined by matching the first two moments, i.e.

$$\begin{cases} \mu = (1-p_0)k\theta + (1-p_0)s \\ \sigma^2 = (1-p_0)(k\theta^2 + p_0(k\theta+s)^2) \end{cases} \Longrightarrow \begin{cases} k = \frac{(\mu - s(1-p_0))^2}{(1-p_0)\sigma^2 - p_0\mu^2} \\ \theta = \frac{(1-p_0)\sigma^2 - p_0\mu^2}{(1-p_0)(\mu - s(1-p_0))} \end{cases}$$

4.1.2 Variance Estimation for omnibus TFisher

Consider one-sided p-values coming from standard normal, $P_i = 1 - \Phi(Z_i)$, $i = 1, ..., n$, Z_i's are multivariate standard normal with correlation $\mathrm{Cov}(Z_i, Z_j) = \rho_{ij}$. Define

$$Y_i = -2\log\left(\frac{P_i}{\tau_2}\right) I(P_i < \tau_1)$$

The Tfisher statistic is defined as

$$W = \sum_{i=1}^{n} -2\log\left(\frac{P_i}{\tau_2}\right) I(P_i < \tau_1) = \sum_{i=1}^{n} Y_i$$

We are interested in the variance of W:

$$\mathrm{Var}(W) = \sum_{i,j} \mathrm{Cov}(Y_i, Y_j) = \sum_{i,j} \mathrm{Cor}(Y_i, Y_j)\mathrm{Var}(Y) = \mathrm{Var}(Y)\left(\sum_{i,j} \mathrm{Cor}(Y_i, Y_j)\right)$$

The reasons for this transformation are 1) we know exactly

$$\mathrm{Var}(Y) = 4\tau_1(1 + (1-\tau_1)(1 - \log\tau_1 + \log\tau_2)^2)$$

and 2) $\mathrm{Cor}(Y_i, Y_j)$ is easier to approximate.

Instead of calculating variance of TFisher statistic, we make a general extension to omnibus TFisher statistic. TFisher statistic can be easily calculated by specify certain parameters such that all τ_{1k}s are equal and all τ_{2k}s are equal.

Consider multiple TFisher statistics

$$W_k = \sum_{i=1}^{n} -2\log\left(\frac{P_i}{\tau_{2k}}\right) I(P_i < \tau_{1k}) = \sum_{i=1}^{n} Y_{ik} \quad k = 1, ..., K$$

where

$$Y_{ik} = -2\log\left(\frac{P_i}{\tau_{2k}}\right) I(P_i < \tau_{1k}) \quad i = 1, ..., n$$

We want model $(W_1, ..., W_K)$ as multivariate normal. Thus we are interested in the covariance between W's:

$$\mathrm{Cov}(W_l, W_k) = \sum_{i,j} \mathrm{Cov}(Y_{il}, Y_{jk}) = \sum_{i,j} \mathrm{Cor}(Y_{il}, Y_{jk}) \sqrt{\mathrm{Var}(Y_{il})\mathrm{Var}(Y_{jk})}$$

where

$$\mathrm{Var}(Y_{il}) = 4\tau_{1l}(1 + (1-\tau_{1l})(1 - \log\tau_{1l} + \log\tau_{2l})^2)$$
$$\mathrm{Var}(Y_{ik}) = 4\tau_{1k}(1 + (1-\tau_{1k})(1 - \log\tau_{1k} + \log\tau_{2k})^2)$$

Conjecture 3. *Let* $b_k = \Phi^{-1}(1-\tau_{1k})$, $b_l = \Phi^{-1}(1-\tau_{1l})$, *we believe*

$$\mathrm{Cor}(Y_{il}, Y_{jk}) \approx \mathrm{Cor}\left(Z_i^2 I(Z_i > b_l), Z_j^2 I(Z_j > b_k)\right)$$

Conjecture 3 is a more general case of Conjecture 1. Next we will focus on the calculation of Conjecture 3.

The theoretical deduction of $\mathrm{Cor}\left(Z_i^2 I(Z_i > b_l), Z_j^2 I(Z_j > b_k)\right)$ is a natural extension of the soft thresholding case where $b_l = b_k$.

Similarly since Z_i, Z_j are bivariate standard normal random variables with correlation ρ_{ij}, they have the same distribution as (U, V):

$$\begin{aligned} U &\sim N(0,1) \\ V &= \rho U + \sqrt{1-\rho^2} Z, \quad \rho = \rho_{ij} \end{aligned}$$

where Z is a standard normal random variable independent with U. Thus $\mathrm{Cor}\left(Z_i^2 I(Z_i > b_l), Z_j^2 I(Z_j > b_k)\right)$ $\mathrm{Cor}\left(U^2 I(U > b_l), V^2 I(V > b_k)\right)$. Under such transform, the underlying r.v. U and Z are independent.

Next we will focus on the calculation of

$$\begin{aligned} &\mathrm{E}(U^2 I(U > b_l) V^2 I(V > b_k) \\ =&\mathrm{E}\left[U^2(\rho U + \sqrt{1-\rho^2} Z)^2 I(U > b_l) I(\rho U + \sqrt{1-\rho^2} Z > b_k)\right] \\ =&\mathrm{E}\left[\left(\rho^2 U^4 + 2\rho\sqrt{1-\rho^2} U^3 Z + (1-\rho^2) U^2 Z^2\right) I(U > b_l) I(\rho U + \sqrt{1-\rho^2} Z > b_k)\right]. \end{aligned} \tag{1}$$

The integration region is $\{b_l < u < +\infty\} \bigcup \{f_k(U) < z < +\infty\}$, where $f_k(u) = \frac{-\rho u + b_k}{\sqrt{1-\rho^2}}$.

Denote $S = \left(\rho^2 U^4 + 2\rho\sqrt{1-\rho^2} U^3 Z + (1-\rho^2) U^2 Z^2\right)$.

Using Corollary 1 and 2 given in previous, we can calculate covariance term step by step.

$$
\begin{aligned}
&\mathrm{E}SI\left(b_l < U < +\infty\right) I(f_k(U) < Z < +\infty) \\
&= \int_{b_l}^{+\infty} \int_{f_k(u)}^{+\infty} \left(\rho^2 u^4 + 2\rho\sqrt{1-\rho^2} u^3 z + (1-\rho^2) u^2 z^2\right) \phi(z)\phi(u) dz du \\
&= \int_{b_l}^{+\infty} \left(\rho^2 u^4 M_0(u;k) + 2\rho\sqrt{1-\rho^2} u^3 M_1(u;k) + (1-\rho^2) u^2 M_2(u;k)\right) \phi(u) du \\
&= \int_{b_l}^{+\infty} u^2 h(u;k) du
\end{aligned}
\tag{2}
$$

where $h(u;k) \stackrel{def}{=} \left(\rho^2 u^2 M_0(u;k) + 2\rho\sqrt{1-\rho^2} u M_1(u;k) + (1-\rho^2) M_2(u;k)\right) \phi(u)$, $M_n(u;k) = \mathrm{E}Z^n I(f_k(u) < Z < +\infty)$.

We conclude that

$$
\begin{aligned}
&\mathrm{E}(U^2 I(U > b_l) V^2 I(V > b_k)) \\
&= \int_{b_l}^{+\infty} u^2 h(u;k) du
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
\mathrm{E}U^2 I(U > b_l) &= \tau_{1l} + b_l \phi(b_l) \\
\mathrm{E}V^2 I(V > b_k) &= \tau_{1k} + b_k \phi(b_k).
\end{aligned}
\tag{4}
$$

Therefore the covariance

$$
\begin{aligned}
&\mathrm{Cov}(U^2 I(U > b_l), V^2 I(V > b_k)) \\
&= \int_{b_l}^{+\infty} u^2 h(u;k) du - (\tau_{1l} + b_l \phi(b_l))(\tau_{1k} + b_k \phi(b_k)).
\end{aligned}
\tag{5}
$$

To find the correlation, we need to calculate the variance,

$$
\begin{aligned}
\mathrm{Var}(U^2 I(U > b_l)) &= \mathrm{E}U^4 I(U > b_l) - (\mathrm{E}U^2 I(U > b_l))^2 \\
&= M_4(b_l) - (M_2(b_l))^2 \\
\mathrm{Var}(V^2 I(V > b_k)) &= M_4(b_k) - (M_2(b_k))^2.
\end{aligned}
\tag{6}
$$

Finally, with

$$\mathrm{Var}(Y_{il}) = 4\tau_{1l}(1 + (1-\tau_{1l})(1 - \log\tau_{1l} + \log\tau_{2l})^2)$$
$$\mathrm{Var}(Y_{ik}) = 4\tau_{1k}(1 + (1-\tau_{1k})(1 - \log\tau_{1k} + \log\tau_{2k})^2)$$

, we have

$$\begin{aligned}
&\mathrm{Cov}(Y_{il}, Y_{jk}) \\
=&\mathrm{Cor}(Y_{il}, Y_{jk})\sqrt{\mathrm{Var}(Y_{il})\mathrm{Var}(Y_{jk})} \\
\approx&\mathrm{Cor}(U^2 I(U > b_l), V^2 I(V > b_k))\sqrt{\mathrm{Var}(Y_{il})\mathrm{Var}(Y_{jk})} \\
=&\mathrm{Cov}(U^2 I(U > b_l), V^2 I(V > b_k))\sqrt{\frac{\mathrm{Var}(Y_{il})\mathrm{Var}(Y_{jk})}{\mathrm{Var}(U^2 I(U > b_l))\mathrm{Var}(V^2 I(V > b_k))}} \\
=&\left[\int_{b_l}^{+\infty} u^2 h(u;k)du - (\tau_{1l} + b_l\phi(b_l))(\tau_{1k} + b_k\phi(b_k))\right] * \\
&4\sqrt{\frac{\tau_{1l}(1 + (1-\tau_{1l})(1 - \log\tau_{1l} + \log\tau_{2l})^2)\tau_{1k}(1 + (1-\tau_{1k})(1 - \log\tau_{1k} + \log\tau_{2k})^2)}{(M_4(b_l) - (M_2(b_l))^2)(M_4(b_k) - (M_2(b_k))^2)}}.
\end{aligned} \tag{7}$$

4.2 Graphs of Variance calculation of TFisher statistic

In TFisher case, we only compare the simulation result with the theoretical one. We fix τ_2 since τ_2 is actually a scale parameter which does not affect the shape of the fitting line.

Figure 4.1 are τ_1 vs variance with correlation fixed. We can find out the theoretical method works fine when τ_1 is small, but it tend to underestimate variance when τ_1 approaches to 1. This can be explained that we only take Z^2 terms as approximation for Y.

When we fix τ_1 and τ_2 in Figure 4.2 and let polynomial decaying coefficient λ varies, we would find the main departure is caused by τ_1, especially when τ_1 gets larger.

Figure 4.1: Variance calculation of TFisher statistic under equal correlation. The τ_1 parameter is chosen such that the theoretical calculation has large deviation from the sample variance.

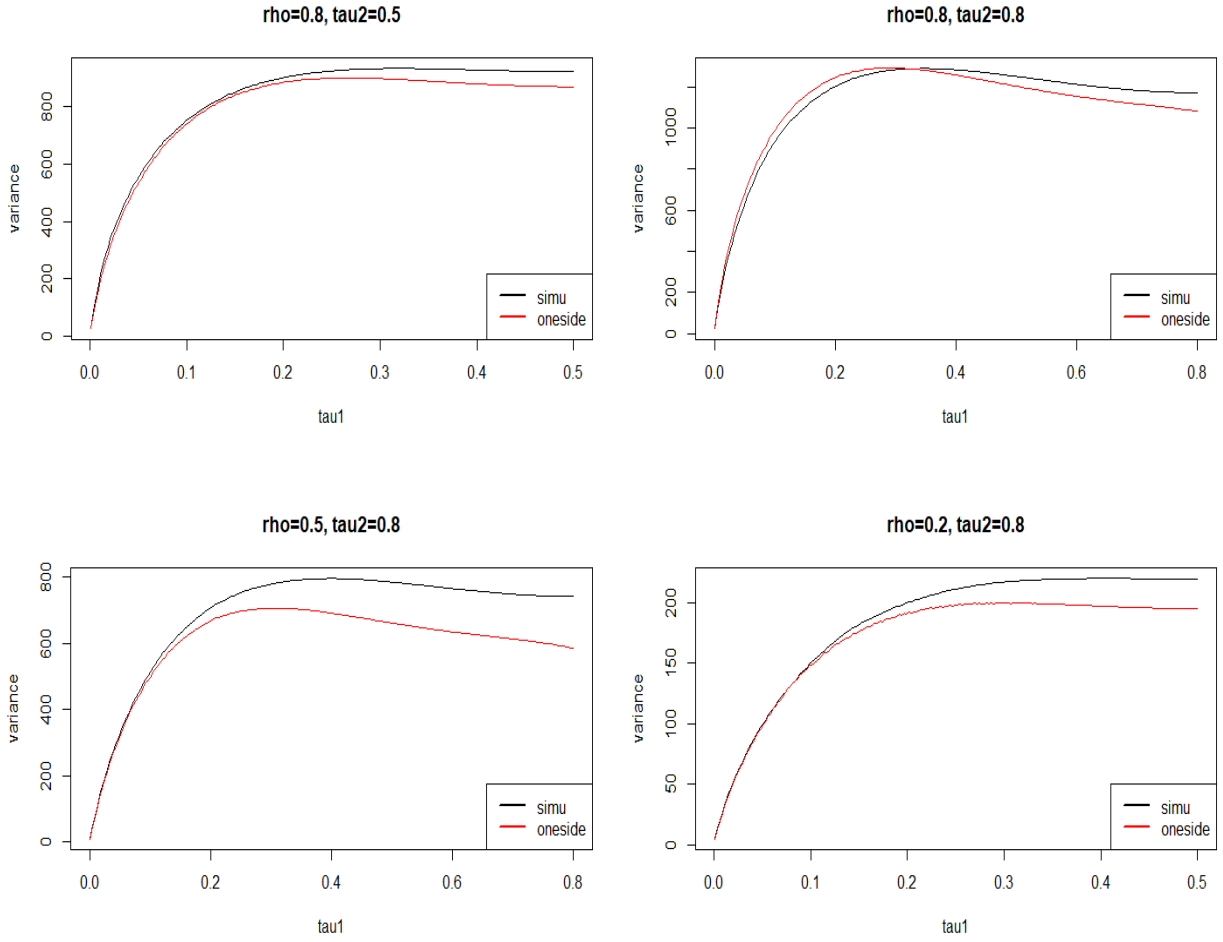
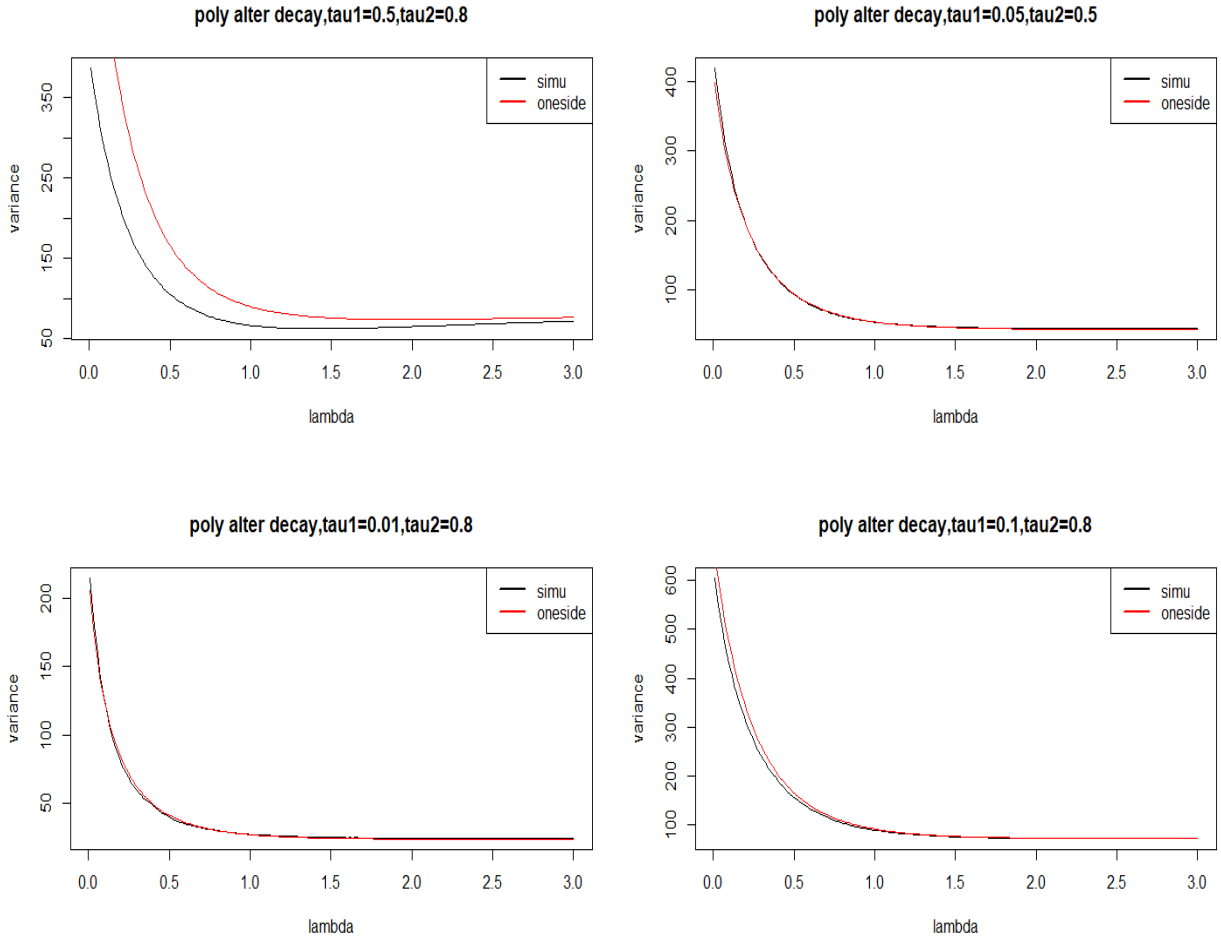


Figure 4.2: Variance calculation of TFisher statistic under polynomial decaying correlation with alternative. The τ_1 parameter is chosen such that the theoretical calculation has large deviation from the sample variance.



4.3 P-value calculation of TFisher statistic

Based on the shifted-mixed Gamma distribution in section 4.1.1, we can calculate p-value by calculating parameters of p_0, θ, k and s.

Shift parameter s is the gap of discontinuity of a general TFisher statistic. It can be

estimated by

$$s = -2\log(frac\tau_1\tau_2)$$

p_0 can be estimated by

$$p_0 = P(P_i > \tau, i = 1, ..., n)$$

which can be calculated by a multivariate normal distribution. k and θ are determined by matching the first two moments, i.e.

$$\begin{cases} \mu = (1-p_0)k\theta + (1-p_0)s \\ \sigma^2 = (1-p_0)(k\theta^2 + p_0(k\theta+s)^2) \end{cases} \Longrightarrow \begin{cases} k = \frac{(\mu - s(1-p_0))^2}{(1-p_0)\sigma^2 - p_0\mu^2} \\ \theta = \frac{(1-p_0)\sigma^2 - p_0\mu^2}{(1-p_0)(\mu - s(1-p_0))} \end{cases}$$

For P-value calculation of TFisher statistic, the first horizontal part explain the p_0 under shifted-mix gamma distribution. An interesting pattern occurs when $\frac{\tau_1}{\tau_2}$ is relatively small($\leq$ 0.1) also with high correlation.

This phenomenon can be explained by the construction of TFisher test statistic W. Recall

$$W = \sum_{i=1}^{n} -2\log\left(\frac{P_i}{\tau_2}\right) I(P_i < \tau_1) = \sum_{i=1}^{n} Y_i$$

Assume τ_1 is small, then if we want $P_j \leq \tau_1$, since$-2log(\frac{\tau_1}{\tau_2}) \leq -2log(\frac{P_j}{\tau_2})$, W would increase at leat $-2log(\frac{\tau_1}{\tau_2})$. Regarding to the Figure 4.3 and 4.4, if $\tau_1 = 0.001$, $\tau_2 = 1$, then for one P_j satisfies $P_j < \tau_1$, W would have to increase $gap = -2log(\frac{\tau_1}{\tau_2}) = 13.8$. When those P_js are highly correlated, then the gap would tend to occur at the value which is close to $n * gap, n = 1, 2, 3, \ldots$.

Figure 4.3: P-value calculation of TFisher statistic under equal correlation.

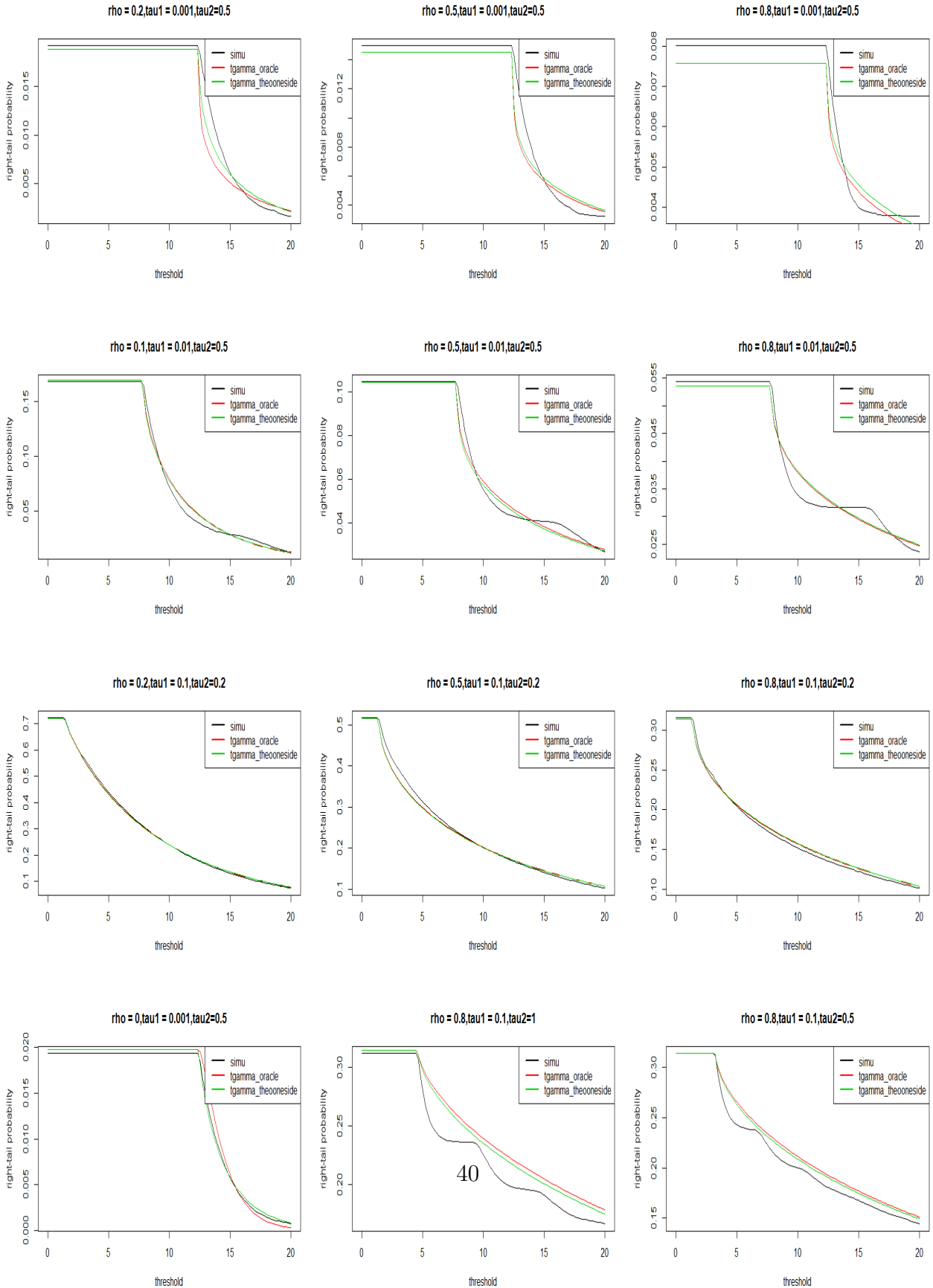
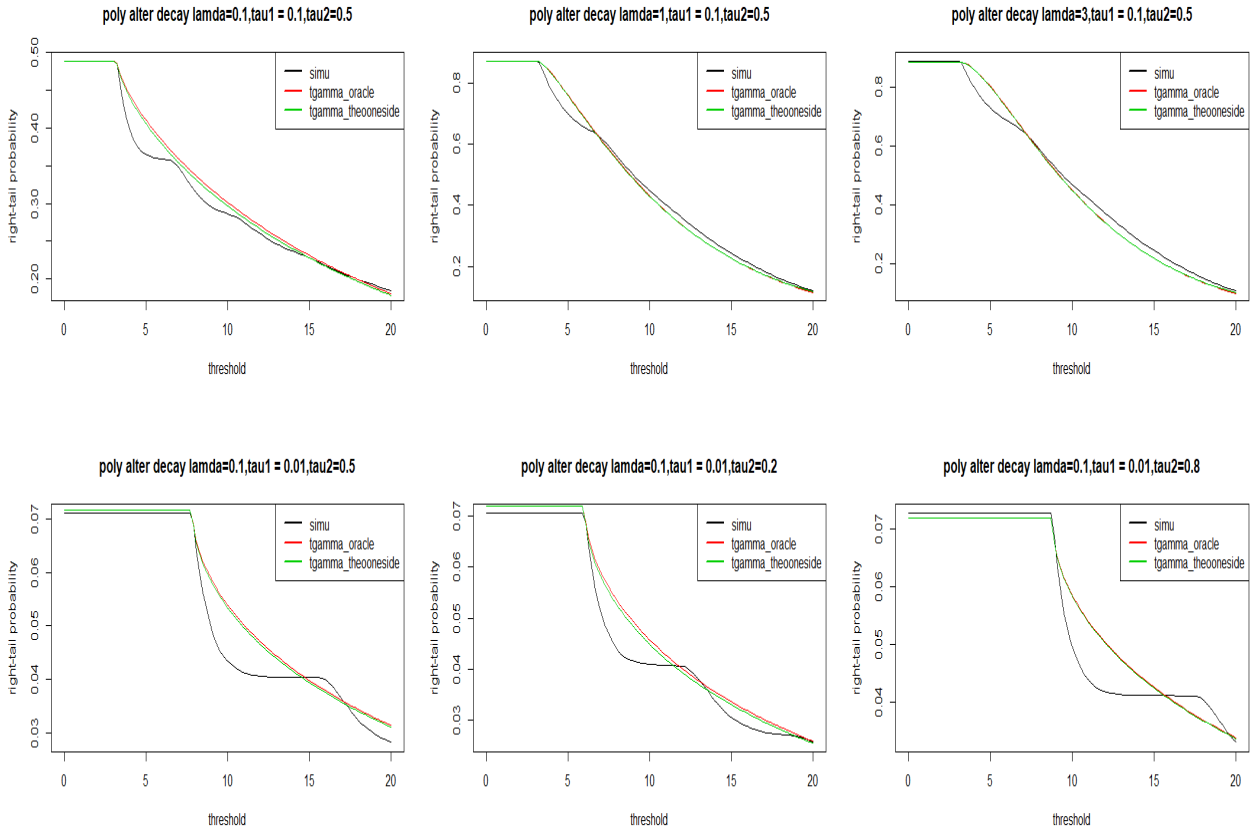


Figure 4.4: P-value calculation of TFisher statistic under polynomial decaying correlation.



Chapter 5

Discussion and Further Improvement

Polynomial fitting would be always an efficient way to find the variance for TFisher statistic, however it cannot guarantee the accuracy when τ(The truncation) is relatively small. Another concern is polynomial has its own shortcoming, especially for the order of correlation part which would cause negative variance by inappropriately chosen.

For the theoretical method of variance calculation, it has been proved to be an improvement comparing to the others. The problem for the theoretical method remain to be the approximation form of Y_i. How we determine the A such that we can get better estimate of variance.

The theoretical method for TFisher statistic seems do not work well in variance calculation and the shift-mixed gamma distribution actually cannot be apply to find out the p-value when there exists high correlation.

The shift-mixed gamma distribution is decent in three parameters cases. One way can possibly improve the performance is add high order moments. Then we would have more information to estimate the variance and further get more precise distribution of W.

Bibliography

[1] Morton B. Brown. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975.

[2] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[3] Zheyang Wu Hong Zhang. Distributions and statistical power of optimal signal-detection methods in finite cases. *arXiv*, 2017.

[4] James T Kost and Michael P McDermott. Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183 – 190, 2002.

[5] Pieter Bastiaan Ober. Asymptotic theory of statistics and probability. *Journal of Applied Statistics*, 38(4):869–869, 2011.

[6] William Poole, David L Gibbs, Ilya Shmulevich, Brady Bernard, and Theo A Knijnenburg. Combining dependent p-values with an empirical adaptation of brownâĂŹs method. *Bioinformatics*, 32(17):i430–i436, 2016.

[7] D.V. Zaykin, Lev A. Zhivotovsky, P.H. Westfall, and B.S. Weir. Truncated product method for combining pâĂŘvalues. *Genetic Epidemiology*, 22(2):170–185, 2002.