Biometrics 64, 1–18 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Fast P-value Calculation for Higher Criticism and Berk-Jones Test in Very Stringent Significance Levels

Wenjia Wang

Department of Biostatistics, University of Pittsburgh

email: wew89@pitt.edu

and

Kathy Author

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

email: anotherauthor@address.edu

Summary:
Key words: Higher criticism; Berk-Jones; Multiple hypothesis testing; Genome-wide association studies; p-value computation.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Traditional Genome-wide association studies (GWASs) identified many single nucleotide polymorphism (SNPs) associated with a large number of human diseases, but the reported common single nucleotide variants (SNVs) often explain only a small proportion of the risk for most diseases, resulting in missing inheritance(Eichler et al., 2010). As the next generation sequencing (NGS) technology developed, all SNVs can be measured and the high-throughput data enabled geneticists to reveal the missing heritability (Mardis, 2008). Among the more massive amounts of SNVs, only a small proportion of them is expected to be phenotype-associated (e.g. associated with a disease of interest), so the traditional multiple hypothesis testing in omics analysis, such as Fisher's p-value combination test (Fisher, 1934), becomes ineffective for detecting the sparse signals and a more powerful test is required in this situation. In addition, since the genetic data generally contains millions of or billions of SNVs and thousands of samples, scientists commonly group multiple SNVs together based on their chromosomal site and test the association between each SNV-set and the phenotype one at a time (Song and Zhang, 2014). This results in tens of thousands of sets and more stringent type I error control for each set test, thus a fast and accurate p-value computation method at stringent significance level is desired (Liu and Xie, 2020).

1.1 *Properties of HC and BJ Statistics*

Many methods of combining individual p-values or test statistics for the multiple hypothesis test were proposed, among which the *higher criticism* (Donoho and Jin, 2004) and the *Berk-Jones* (Berk and Jones, 1979) are particularly famous for their improved power for detecting weak and sparse alternatives.

The individual test statistics are denoted by $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_K})^T$ where K is the dimension of the set, and we assume that $\{\mathbf{x_i}\}_{i=1}^{K}$ independently and identically follow standard normal distribution under null, while a small proportion of ϵ_K has mean shift

μ_K under alternative:

$$H_0 : \mathbf{x_i} \overset{i.i.d.}{\sim} N(0,1) \quad vs. \quad H_1 : \mathbf{x_i} \overset{i.i.d.}{\sim} (1-\epsilon_K)N(0,1) + \epsilon_K N(\mu_K, 1). \tag{1}$$

The *higher criticism statistic* (HC) proposed by Donoho and Jin (2004) aggregates multiple effects by combining individual one-sided or two-sided p-values which are $p_i = 1 - \Phi(x_i)$ or $p_i = 2(1 - \Phi(\|x_i\|))$ respectively for $i = 1, \ldots, K$, where Φ is the CDF of standard normal distribution and x_i is the realization of $\mathbf{x_i}$. The original HC test statistic is given by

$$T_{HC^+} = \max_{1 \leqslant i \leqslant \alpha_0 K} \sqrt{K} \frac{i/K - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}, \tag{2}$$

where $p_{(1)}, ..., p_{(K)}$ are ordered p-values $\{p_i\}_{i=1}^K$ and $\alpha_0 \in (0, 1]$. Donoho and Jin (2004) also demonstrated that it is asymptotically optimal in the *detection boundary* sense under the setting of weak and sparse signals. Cai and Wu (2014) extended the alternative distribution from the mixed Gaussian to a more general mixture form, then provided an asymptotic detection threshold in this mixture setting and proved the asymptotic optimality of T_{HC^+}. For practical application, in addition to using the *full domain HC* (T_{FHC}) by setting $\alpha_0 = 1$ in T_{HC^+}, *half domain HC* (T_{HHC}) is also suggested with the value of $\alpha_0 = 0.5$ (Donoho and Jin, 2004). Furthermore, when the signals are severely weak, the modified version of HC with supremum domain $\{1 \leqslant i \leqslant K/2, p_{(i)} > 1/K\}$, which actually excludes a few smallest individual p-values for more power gains, is preferable in finite samples (Donoho and Jin, 2004). Therefore, a generalized form of HC with flexible domain is

$$T_{HC} = \max_{k_0 \leqslant i \leqslant k_1} \sqrt{K} \frac{i/K - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}, \tag{3}$$

where $1 \leqslant k_0 \leqslant k_1 \leqslant K$ (Zhang et al., 2020; Li and Siegmund, 2015).

Another appealing class of goodness-of-fit (GOF) test statistics for detecting weak and sparse signals is proposed by Berk and Jones (1979), and the one-sided version corresponding to the same scenario of the HC, called *the Berk-Jones statistic* (BJ), is in the following

formula:

$$T_{BJ} = \max_{1 \leqslant i \leqslant K} \mathbb{1}\{p_{(i)} < \frac{i}{K}\} \sqrt{2K[\frac{i}{K} log(\frac{i/K}{p_{(i)}}) + (1 - \frac{i}{K}) log(\frac{1 - i/K}{1 - p_{(i)}})]}. \tag{4}$$

BJ is asymptotically optimal for detecting weak and sparse signals in the setting of mixed Gaussian alternatives, and it shares the same asymptotic *detection boundary* with T_{HC^+} (Donoho and Jin, 2004). Moreover, under some alternatives, BJ may have considerably higher power compared to HC in finite samples (Moscovich et al., 2016).

In order to use HC or BJ to make decision on the hypothesis test, we also have to figure out the distribution of T_{HC} (3) and T_{BJ} (4) under null hypothesis, which is either for p-value computation given the observed value of the test statistic or finding the threshold given the significance level. Denote the survival function for T_{HC} and T_{BJ} under null hypothesis are $\bar{F}_{HC}(q) = \mathbb{P}(T_{HC} > q|H_0)$ and $\bar{F}_{BJ}(q) = \mathbb{P}(T_{BJ} > q|H_0)$ respectively. Some methods were proposed in literature to reveal $\bar{F}_{HC}(q)$ and $\bar{F}_{BJ}(q)$, but they have various limitations.

1.2 *Limitations of the Current Methods for Computing $\bar{F}_{HC}(q)$ and $\bar{F}_{BJ}(q)$*

In general, there are three classes of methods to calculate the survival functions of T_{HC} and T_{BJ} under the null hypothesis. The first is the naive Monte-Carlo simulation or permutation tests. These methods are flexible for all test statistic including T_{BJ} and T_{HC} with any supremum domain, and theoretically unbiased and efficient as long as the sample size is sufficiently large. However, one of the most significant limitations of Monte Carlo is that it is unable to calculate the extremely right tail probability i.e. $\bar{F}_{HC}(q)$ or $\bar{F}_{HC}(q)$ for very large q, or it manages to reveal the right tail of the null distribution but at the cost of intensive computation. Therefore, the class of methods becomes infeasible when the right tail of the null distribution really matters.

The second class of methods is to approximate the null distribution by asymptotic theory. However, one problem is that even though both T_{HC} and T_{BJ} follow a known distribution asymptotically under the null, the convergence is very slow. Thus except for a large dimension

K, the asymptotic null distribution is unreliable (Barnett and Lin, 2014; Jaeschke, 1979). Furthermore, some asymptotic approximation methods either fail to approximate the whole null distribution or cannot cover arbitrary supremum domain of T_{HC} (Li and Siegmund, 2015; Zhang et al., 2020).

The last type of methods is to analytically calculate the exact null distribution of T_{HC} and T_{BJ} (Khmaladze and Shinjikashvili, 2001; Noe, 1972). Barnett and Lin (2014) provided an analytic method (denoted by LIN) specifically for *full domain HC*. These recursive methods indeed can can obtain theoretically accurate null distribution, but they require daunting computation with complexity of $O(K^3)$ for high dimension in particular. Though the computation method in Moscovich et al. (2016) reduced the complexity to $O(K^2)$, it is only applicable for calculating the null distribution of *full domain HC* (T_{FHC}) and T_{BJ}. Zhang et al. (2020) developed an R package called "SetTest" which can be applied to generalized HC (T_{HC}) and T_{BJ} with computational complexity of $O(K^2)$, but the resulting null distribution suffers from numerical unstability at the extremely right tail. What's worse, the implementation of "SetTest" contains the numerical approximation which significantly affects the accuracy of the null distribution for T_{BJ}. Moreover, naively removing the numerical approximation step will result in a fast accumulation of numerical errors, and consequently the algorithm will break down completely at $K \approx 150$ (Moscovich et al., 2016).

[Table 1 about here.]

[Table 2 about here.]

1.3 *Our Contributions*

Our main contributions are (1) suggesting the cross-entropy-based importance sampling method (IS) as the benchmark to assess the method LIN and SetTest, especially in terms of their accuracy in computing the very right tail of the null distribution for T_{HC}; (2) proposing the improved methods (HC.mod) for computing T_{HC} null distribution by modifying SetTest,

and embedding the translated polynomial technique (TP) in calculating the exact null distribution of T_{BJ} to reduce the accumulation of numerical error; (3) using our proposed methods to construct the quantile-probability tables and fit smooth curves, subsequently to reveal the whole null distribution of T_{HC} and T_{BJ} with various dimensions K. These smooth curves in fact provide a new approach for vectorizing computation of p-values given a huge number of observed value of T_{HC} or T_{BJ}, or finding the thresholds very fast given a family-wise significance level.

In section 2, we propose the cross-entropy-based importance sampling method (IS), the modified method (HC.mod) for computing T_{HC} null distribution and embed the translated polynomial technique in computing T_{BJ} null distribution. In section 3, we justify our proposed methods and compare them with LIN and SetTest in terms of accuracy and time complexity based on simulation. In section 4, we use our proposed methods to construct the quantile-probability libraries for T_{FHC}, T_{HHC} and T_{BJ} of dimension K ranging from 1 to 2000 and right-tail probability small to 10^{-14}. Then, we wrap up the libraries and develop the R package "STlibs", then conduct simulation comparison to demonstrate that STlibs is faster and accurate enough in coding with a large number of HC or BJ tests. We show the applications of STlibs in a real GWAS in Section 5.

2. Our Methods

One of our purposes is to find a fast and accurate approach to compute the null distribution for T_{HC} and T_{BJ} with various dimension K. Since we target at the whole null distribution, the performance of the algorithm in revealing the tail part matters. Current literature (Zhang et al., 2020; Li and Siegmund, 2015; Barnett and Lin, 2014; Moscovich et al., 2016) has rarely discussed the efficiency of their methods in computing extremely right tail probability (i.e. $\bar{F}_{HC}(q)$ or $\bar{F}_{BJ}(q)$ at large value of q). Therefore, in this section, we propose two methods (IS and HC.mod) of computing the null distribution for T_{HC} with flexible supremum domain

and one method (TP) for T_{BJ}, which are superior to the current methods (e.g. LIN and SetTest) in revealing the right tail of the null distribution of moderate or large dimension K in terms of numerical accuracy and time complexity.

2.1 *Importance Sampling Method of Computing the Null Distribution of T_{HC}*

Monte Carlo sampling method is a popular way to approximate the expected value of a random quantity. However, for T_{HC} whose null distribution is heavy-tailed, the naive Monte Carlo is less efficient for computing the survival function $\bar{F}_{HC}(q)$ at large value of q unless we significantly increase sample size, which will instead incur demanding computation. In order to improve the sampling efficiency and control the variance of the survival function estimates with a feasible sample size, one method that was proven useful in many settings is the *importance sampling technique* (IS). The key of this technique is independently sampling from an alternative density called *IS density*, which actually increases the probability of drawing large values from the tail, and consequently the samples can better mimic the tail of the null distribution. In our case, we sample $X^{(1)}, X^{(2)}, \ldots, X^{(N)}$, where $X^{(j)} = (x_1^{(j)}, \ldots, x_K^{(j)})^T \in \mathbb{R}^{\mathbb{K}}$ and $\{x_i^{(j)}\}_{i=1}^K$ are independently from the IS density $\{g_i(x)\}_{i=1}^K$ instead of the null distribution $N(0,1)$. The estimated tail probability of T_{HC} under the null at the quantile q (i.e. $\bar{F}_{HC}(q)$) by IS technique is

$$\hat{p}_{IS}(q) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{T_{HC}(X^{(j)}) > q\} \frac{\prod_{i=1}^K \phi(x_i^{(j)})}{\prod_{i=1}^K g_i(x_i^{(j)})}, \tag{5}$$

where $\phi(x)$ is the pdf of $N(0,1)$. The *IS density* $g_i(x)$ is required to dominate the null density in the absolutely continuous sense, and the choice of the *IS density* is of vital importance.

A common approach to choose an appropriate *IS density* is restricting it to some parametric family, and then find the best value of the parameter based on some criteria (e.g. minimizing the variance of the resulting estimates) or by using adaptive procedure to estimate the optimal parameter from a specific form (Oh and Berger, 1992; Vázquez-Abad and Dufresne, 1998). Theoretically, the optimal *IS density* corresponding to zero-variance estimator $\hat{p}_{IS}(q)$

should be $g_{opt}(\mathbf{x}, q) = \mathbb{1}\{T_{HC}(\mathbf{x}) > q\}/p$, where $p = \bar{F}_{HC}(q)$ is the ture tail probability at q. However, this zero-variance density is not applicable since the true $\bar{F}_{HC}(q)$ is what we want to estimate and unknown. One of the approaches to optimize the parameter is the *cross-entropy method* (CE) proposed by Rubinstein (2002), which is to find the parameter value given a parametric family such that the *IS density* with this parameter can minimize the Kullback-Leibler cross entropy with respect to theoretically optimal *IS density* $g_{opt}(\mathbf{x}, q)$.

In the literature, the idea of *importance sampling* has already been used in p-value computation of multiple hypothesis test. For example, Huo et al. (2019) proposed a beta-distribution density as the *IS density*. However, both the choice of parametric family and the best value of the parameter are mainly based on prior knowledge or just randomly picked, thus not universally efficient. Considering the mixed normal alternative of our hypothesis setting, we suggest selecting the best *IS density* from a mixed normal parametric family with parameter θ:

$$g_i(x;\theta) \sim \frac{1}{i+1}N(0,\theta) + \frac{i}{i+1}N(0,1). \tag{6}$$

Then, given any quantile q, we used an iterative procedure to determine the best choice of $\theta(q)$. The following steps adaptively update $\theta(q)$ based on the *cross-entropy* and use the final choice $\theta_{opt}(q)$ to obtain the combined p-value estimate $\hat{p}_{IS}(q)$ (Rubinstein, 2002):

- Step 0: Initial value $\theta^{(0)}(q) = 1$;
- Step 1: Draw N samples $X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(N)}$ from density $g_i(x;\theta^{(t)}(q)) \sim \frac{1}{i+1}N(0,\theta^{(t)}(q)) + \frac{i}{i+1}N(0,1)$ where $X_t^{(j)} = (x_{t,1}^{(j)}, \ldots, x_{t,K}^{(j)})$, i.e. $x_{t,i}^{(j)} \overset{i.i.d.}{\sim} \frac{1}{i+1}N(0,\theta^{(t)}(q)) + \frac{i}{i+1}N(0,1)$ for $j = 1, \ldots, N; i = 1, \ldots, K$, and calculate the corresponding combined statistic $z_t^{(j)} = T_{HC}(X_t^{(j)})$ for $j = 1, \ldots, N$;
- Step2: Compute the empirical 0.99 quantile of $\{z_t^{(j)}\}_{j=1}^N$, and define the temporary threshold $q^{(t)}$ is the minimum of the empirical 0.99 quantile and the target quantile q.

- Step 3: Update the parameter by
$\theta^{(t+1)}(q) = argmax_{\theta(q)}\{\frac{1}{N}\sum_{j=1}^{N}[\mathbb{1}\{T_{HC}(X_t^{(j)}) > q^{(t)}\}\frac{\prod_{i=1}^{K}\phi(x_{t,i}^{(j)})}{\prod_{i=1}^{K}g_i(x_{t,i}^{(j)};\theta^{(t)}(q))} * \log[\prod_{i=1}^{K}g_i(x_{t,i}^{(j)};\theta(q))]]\}$;
- Step 4: If the temporary threshold $q^{(t)}$ is smaller than the target q, repeat step 1-3 until $q^{(t)} \geqslant q$, and then obtain the final parameter $\theta_{opt}(q)$;
- Step 4: Draw N samples $X^{(1)}, \ldots, X^{(N)}$ from the final selected *IS density* $\{g_i(x;\theta_{opt}(q))\}_{i=1}^{N}$, and estimate $\bar{F}_{HC}(q)$ by $\hat{p}_{IS}(q) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{T_{HC}(X^{(j)}) > q\}\frac{\prod_{i=1}^{K}\phi(x_i^{(j)})}{\prod_{i=1}^{K}g_i(x_i^{(j)};\theta_{opt}(q))}$.

Theoretically, the IS estimator, similar to naive Monte Carlo, is unbiased with variance decreasing as the sample size N increases. Attractively, it is superior to naive Monte Carlo in computing small combined p-value. By comparison with other methods, our numerical results in Section 3 will illustrate the efficiency of IS method in particular when computing the tail probability of the null distribution for T_{HC} with flexible supremum domain and various dimension K.

2.2 *Modified Analytic Method of Computing the Null distribution of* T_{HC}

As we discussed before, the limitations of the current analytic methods of computing $\bar{F}_{HC}(q)$ include intensive computation of order $O(K^3)$, restricted supremum domain (e.g. for LIN method), or numerical unstability when q is extremely large under moderate or high dimension K (e.g. for SetTest). Thus, in this part, we will propose an analytic method (HC.mod) by modifying SetTest to precisely calculate the null distribution of T_{HC}, even for extremely right tail part, under various dimensions with any supremum domain at the order $O(K^2)$.

In fact, the null distribution of T_{HC} can be converted into the distribution of order statistics $\{p_{(i)}\}_{i=k_0}^{k_1}$, where $\{p_i\}_{i=1}^{K} \overset{i.i.d.}{\sim} Uniform(0,1)$. SetTest derives the CDF $(1-\bar{F}_{HC}(q))$ by the high-order integral:

$$\begin{aligned}
\mathbb{P}(T_{HC} \leqslant q) &= \mathbb{P}\{p_{(i)} \geqslant L_i, k_0 \leqslant i \leqslant k_1\} \\
&= \int_{L_{k_1}}^{1} \frac{K!}{(K-k_1)!}(1-x_{k_1})^{K-k_1} \int_{L_{k_1-1}}^{x_{k_1}} \cdots \int_{L_{k_0}}^{x_{k_0+1}} \frac{x_{k_0}^{k_0-1}}{(k_0-1)!}\, dx_{k_0} \ldots dx_{k_1} \\
&= \int_{L_{k_1}}^{1} \frac{K!}{(K-k_1)!}(1-x_{k_1})^{K-k_1} \int_{L_{k_1-1}}^{x_{k_1}} \cdots \int_{L_{k_0+1}}^{x_{k_0+2}} \frac{x_{k_0+1}^{k_0}}{k_0!}\, dx_{k_0+1} \ldots dx_{k_1} - \frac{L_{k_0}^{k_0}}{k_0!} a_{k_0+1} \\
&= \bar{F}_{B(k_1,K-k_1+1)}(L_{k_1}) - \sum_{i=k_0}^{k_1-1} \frac{L_i^i}{i!} a_{i+1}, \qquad (7)
\end{aligned}$$

where $\bar{F}_{B(\alpha,\beta)}(x)$ denotes the survival function of $Beta(\alpha, \beta)$, L_i is the root of $f_i(x)$. The functions $\{f_i(x)\}_{i=k_0}^{k_1}$ and $\{a_i\}_{i=k_0+1}^{k_1}$ are defined by

$$\begin{aligned}
f_i(x) &= \sqrt{K}\frac{i/K - x}{\sqrt{x(1-x)}} - q, \\
a_{k_1} &= \frac{K!}{(K-k_1+1)!}\bar{F}_{B(1,K-k_1+1)}(L_{k_1}), \\
a_i &= \int_{L_{k_1}}^{1} \frac{K!}{(K-k_1)!}(1-x_{k_1})^{K-k_1} \int_{L_{k_1-1}}^{x_{k_1}} \cdots \int_{L_i}^{x_{i+1}} dx_i \ldots dx_{k_1} \\
&= \frac{K!}{(K-i+1)!}\bar{F}_{B(k_1-i+1,K-k_1+1)}(L_{k_1}) - \sum_{j=1}^{k_1-i} \frac{L_{i+j-1}^{j}}{j!} a_{i+j}, \quad i = k_0+1, \ldots, k_1-1. \qquad (8)
\end{aligned}$$

Though Zhang et al. (2020) provided the theoretically exact formula, SetTest actually involves a numerical parameter t with default value of 30, which means that when $k_1 - k_0 + 1 > t$, it approximates a_i by

$$\tilde{a_i} = \int_{L_{k_1}}^{1} \frac{K!}{(K-k_1)!}(1-x_{k_0+t-1})^{K-k_1} \int_{L_{k_0+t-2}}^{x_{k_0+t-1}} \cdots \int_{L_i}^{x_{i+1}} dx_i \ldots dx_{k_0+t-1}. \qquad (9)$$

In addition, SetTest restricts the root searching interval to $(10^{-15}, i/K)$ for L_i, which means that the roots will be zero if they are smaller than 10^{-15}. Furthermore, given any target quantile q, SetTest actually first computes the probabilities $1 - \bar{F}_{HC}(q_1), \ldots, 1 - \bar{F}_{HC}(q_5)$, where $q_1, \ldots, q_5$ are 5 points in the interval $[0.9q, 1.1q]$ with equal space. Then it implements a local smooth fitting, and finally obtains the probability $1 - \bar{F}_{HC}(q)$ by interpolation.

The HC.mod method we propose includes four main modifications of SetTest to successfully reduce the numerical error:

- Modification 1: conduct the local smooth curve fitting at logarithm scale instead of at the original scale;
- Modification 2: change the root searching interval from $(10^{-15}, i/K)$ to $(10^{-18}, i/K)$ to obtain more precise roots $\{L_i\}_{i=k_0}^{k_1}$;
- Modification 3: remove the numerical truncation parameter t to get rid of the numerical approximation;
- Modification 4: Do exp-log transformation in computing $\{a_i\}_{i=k_0+1}^{k_1}$ to avoid infinity.

The improvement of the HC.mod method is obvious in our simulation comparison, and the necessity of each modification step is also justified by simulation in Section 3.

2.3 *Modified Analytic Method of Computing the Null distribution of T_{BJ}*

Deriving CDF for T_{BJ} is similar to the formula for T_{HC}, but replace the root functions $\{f_i(x)\}$ in equation 8 with the following functions,

$$f_i(x) = \sqrt{2K[\frac{i}{K}log(\frac{i/K}{p_{(i)}}) + (1 - \frac{i}{K})log(\frac{1 - i/K}{1 - p_{(i)}})]} - q, \tag{10}$$

where $x \in (0, i/K)$. SetTest follows the same procedure as T_{HC} to calculate the CDF of the null distribution for T_{BJ} (i.e. $1 - \bar{F}_{BJ}(q)$), but in this case, the numerical approximation as equation 9 severely damages the accuracy. The four modifications proposed above even fail to remedy the fast error accumulation at high dimension K. From the recursively generating $a_{k1}, \ldots a_{k_0+1}$ in equation 8, we can see that each a_i accumulates the errors from all previous term $\{a_j\}_{j=i+1}^{k_1}$. These errors propagate to the a_{i-1} and are repeatedly amplified, and eventually result in severe error accumulation to a_{k_0+1}. This can also explain why SetTest will break down for T_{BJ} of moderate or large dimension K if we just remove its numerical truncation approximation (i.e. calculate the exact Kth-degree instead of its truncated tth-degreer integral where t=30 by default).

Moscovich et al. (2016) used the translated polynomials technique to reduce the numerical errors when computing the p-value of the exact Berk-Jones test statistic (different from T_{BJ}

here). Thus, in order to attenuate the accumulation of numerical errors, we borrow the idea of translated polynomials (TP) and implement this TP method in R to calculate the null distribution of our T_{BJ} of moderate or high dimension. Since the T_{BJ} is the maximum over full domain $1 \leqslant i \leqslant K$, the equation 7 changes to a Kth-order integral:

$$\begin{aligned}\mathbb{P}(T_{BJ} \leqslant q) &= \mathbb{P}\{p_{(i)} \geqslant L_i, 1 \leqslant i \leqslant K\} \\ &= K! \int_{L_K}^{1} dx_K \int_{L_{K-1}}^{x_K} dx_{K-1} \cdots \int_{L_1}^{x_2} dx_1 \\ &= K! h_K(1), \end{aligned} \tag{11}$$

where we define the ith-degree polynomial $\{h_i(t)\}_{i=1}^K$ by $h_0(t) = 1, h_1(t) = \int_{L_1}^t h_0(x)\,dx, \ldots,$ $h_K(t) = \int_{L_K}^t h_{K-1}(x)\,dx$, and $\{L_i\}_{i=1}^K$ are the roots of corresponding function $\{f_i(x)\}_{i=1}^K$ in equation 10. In order to calculate the integral of equation 11, we need to recursively evaluate $\{h_i(t)\}_{i=1}^K$ to finally obtain $h_K(t)$. Instead of the standard basis for polynomials where $h_i(t) = \sum_{d=0}^i c_{d,i} t^d$, translated polynomials technique (TP) uses a basis of translated monomials for polynomial $h_i(t)$ where $h_i(t)$ can be expressed as $c_{0,i} + \sum_{d=1}^i c_{d,i}(t + z_{d,i})^d$ to update the coefficients $\{c_{d,i}\}_{d=0}^i$ and $\{z_{d,i}\}_{d=1}^i$ in each iteration $i = 1, \ldots, K$. With this translated basis, deriving $h_i(t) = \int_{L_i}^t h_{i-1}(x)\,dx$ yields equation 12,

$$\begin{aligned} h_i(t) &= \int_{L_i}^{t} h_{i-1}(x)\,dx = c_{0,i-1}(t - L_i) \\ &\quad + \sum_{d=2}^{i} \frac{c_{d-1,i-1}}{d}(t + z_{d-1,i-1})^d - \sum_{d=2}^{i} \frac{c_{d-1,i-1}}{d}(L_i + z_{d-1,i-1})^d \\ &= c_{0,i} + \sum_{d=1}^{i} c_{d,i}(t + z_{d,i})^d, \end{aligned} \tag{12}$$

and consequently, the coefficients can be updated by their relationship:

$$\begin{aligned} &z_{1,i} = -L_i, \quad z_{d,i} = z_{d-1,i-1}, \quad d = 2, \ldots, i; \\ &c_{0,i} = -\sum_{d=2}^{i} \frac{c_{d-1,i-1}}{d}(L_i + z_{d-1,i-1})^d, \quad c_{d,i} = \frac{c_{d-1,i-1}}{d}, \quad d = 1, \ldots, i. \end{aligned} \tag{13}$$

The updating rule as equation 13 updates $c_{0,i}$ only by the previous term $\{c_{d,i-1}\}_{d=1}^{i-1}$ without $c_{0,i-1}$, so theoretically the accumulation of numerical error in getting the final coefficients

$c_{0,K}, c_{1,K}, \ldots, c_{K,K}$ of $h_K(t)$ will be alleviated. The simulation in Section 3 illustrates the accuracy gains, compared to SetTest. Theoretically, the computation burden is in quadratic order($O(K^2)$), and to further reduce the numerical errors in high dimension, our TP method also uses extended double-precision (80-bit) to compute the null distribution for T_{BJ} of dimension $K > 150$. Although this stringent requirement of precision slows down the computation to some degree, the time complexity shown in Section 3 is acceptable.

3. Methods Evaluation

In this section, we evaluate our proposed methods (IS, HC.mod and TP) and the current methods (Lin, SetTest and LS) in terms of the accuracy and time complexity by simulation experiments and eventually, based on the simulation results, determine the most efficient computation methods of revealing the null distribution of T_{HC} with flexible supremum domain and T_{BJ} under various dimensions to construct our quantile-probability libraries.

3.1 *Evaluating Methods for T_{HC}*

In this part, we start with showing the issues of the two prevalent analytic methods LIN and SetTest about computing the null distribution of T_{FHC} by comparing them with IS. Considering the main concern in computing the extremely right tail probability, we use different methods to compute $\bar{F}_{HC}(q)$ corresponding to 200 quantiles q (i.e. p-value at the given observed statistic q), where the target range of $\bar{F}_{HC}(q)$ is from 10^{-14} to 10^{-6} to reveal the right tail of the null distribution, and compare the results. In addition, We pick a moderate dimension $K = 100$ and a relatively high dimension $K = 500$ for the *full domain HC* (T_{FHC}). About the importance sampling method, we always use sample size of $N = 10^4$, and provide the coefficient of variation based on 10 replications in all simulations.

Figure 1 first compares the LIN and SetTest (column 1) for T_{FHC} under dimension $K = 100$ (upper) and $K = 500$ (bottom). We can see that the resulting right tail probabilities (p-

values) by LIN and SetTest are no longer the same when the significance level smaller than 10^{-13} under $K = 100$ and 10^{-12} under $K = 500$ respectively. Since both are analytic methods, this may be a clue for us to suspect that the accumulation of numerical errors may become a severe problem, leading to the departure. Then, comparing LIN and SetTest with IS respectively (column 2 and 3 in Figure 1) implies that LIN is sufficiently accurate while SetTest is severely problematic in calculating extremely right tail probability. In addition, the coefficient of variation (the error bar in plots) of the probability estimates by IS can be controlled under 0.6 for both dimensions, even when estimating $\bar{F}_{HC}(q)$ smaller than 10^{-14}, which evidences that IS method is a good rescue to estimate $\bar{F}_{HC}(q)$ at extremely large q where naive Monte Carlo definitely fails. The coincident results between IS and HC.mod in Figure 1 (column 4) justify the improvement of the accuracy by HC.mod method.

We conduct the same method comparisons for T_{HHC}, but since LIN is restricted to T_{FHC}, only SetTest, HC.mod and IS methods are compared. Figure 2 manifests that SetTest suffers from the same numerical difficulties in revealing the right tail of the null distribution for T_{HHC} (column 1), while the HC.mod and IS are still similarly accurate and stable in computing the whole null distribution under both dimensions (column 2).

In addition, the Supplementary Figure 4 and Figure 5 justify the necessity of the four modification steps in precisely and stably computing the null distribution of T_{FHC} and T_{HHC}. The simulation settings are the same as before, except we only use dimension of $K = 500$, considering the numerical defeat of SetTest is much worse under high dimension. It turns out that the first two modifications are critical to improve the numerical stability for both T_{FHC} and T_{HHC}, while the numerical approximation $t = 30$ by default in SetTest does not cause significant loss of accuracy. However, removal of the numerical approximation must be combined with involving the exp-log transformation of Modification 4, otherwise it

would break down. Therefore, for both T_{FHC} and T_{HHC}, embedding Modification 1 and 2 is sufficient to obtain the precise right tail probability.

3.2 *Evaluating Methods for* T_{BJ}

In order to evaluate our TP method of computing the null distribution for T_{BJ}, we include the approximation method proposed by Li and Siegmund (2015) (denoted by LS) in our simulation comparison. LS method asymptotically approximates the null distribution of T_{BJ}, and theoretically it is inadequate to compute $\bar{F}_{BJ}(q)$ at small value of q but the approximation is sufficiently accurate to reveal the very right tail distribution for high dimension T_{BJ} (Zhang et al., 2020). Thus, LS is a good choice to facilitate the assessment of SetTest and TP at stringent significance level. Figure 3 compares the probabilities $\bar{F}_{BJ}(q)$ generated by SetTest, TP and LS methods given 200 values of q under three dimensions of T_{BJ} ($K = 30, 100$ and 500) where the $\bar{F}_{BJ}(q)$ ranging from 10^{-13} to 10^{-3}, and also illustrates LS is not precise enough in calculating large probability (larger than 0.01 with $K = 30$).

As the first column in Figure 3 shows, when the dimension $K = 30$, SetTest does not trigger the numerical approximation (since $t = 30$ by default), so the resulting null distribution of T_{BJ} revealed by SetTest, LS and TP almost exactly coincides. When $K > 30$, as the second and third columns in Figure 3 illustrates, the results by LS and TP are still matched but SetTest is away from them across the whole range of the null distribution. Moreover, with the dimension increasing, SetTest becomes less and less precise since the numerical truncation $t = 30$ reduces the Kth-degree integral to 30th-degree integral, resulting in the probabilities departing from the truth. The performance of LS (the rightest plot in Figure 3) is consistent with what we expect that the asymptotic approximation of the null distribution under moderate or small dimension is not sufficiently precise at high significance level, thus failing to reveal the whole null distribution.

3.3 *Comparison of Time Consuming*

From the perspective of time complexity, Table 3 compares the time consuming of Lin, SetTest, IS and HC.mod for calculating the probabilities $\bar{F}_{HC}(q)$ of T_{FHC} and T_{HHC} given 200 values of q. SetTest and HC.mod are competitively fastest, while IS is acceptably slower but it is faster than LIN under high dimension. Thus, though accurate LIN is, the heavy computation burden ($O(K^3)$) especially for large K is a significant barrier.

Regarding the computation methods for T_{BJ}, the time used for SetTest, TP and LS to calculate the probabilities $\bar{F}_{BJ}(q)$ given 200 values of q is displayed in Table 3. Obviously, the approximation LS is fastest, while TP slightly slows down under high dimension due to using the extended precision (80-bit) floating point numbers to attenuate error accumulation. All the computation uses 30 cores in parallel in 32 cores / 64 threads (2 x Sixteen-Core Intel Xeon Gold 6130 2.1G) processor.

In conclusion, HC.mod is applicable for T_{HC} with flexible supremum domain and outperforms the other computation methods in precisely revealing the extremely right tail of the null distribution for T_{HC} of moderate or large dimensions without any cost of computation time. Thus, HC.mod is a good approach to efficiently construct the quantile-probability libraries for the whole null distribution of T_{HC} with flexible domain and various dimensions. The libraries for the whole null distribution for T_{BJ} of low dimensions ($K \leqslant 150$) will be generated by TP method, considering its advantages of accuracy and fast computation in low dimension scenario. However, for T_{BJ} of high dimensions ($K > 150$), the combination of LS and TP methods is a better choice, which not only retains the accuracy of TP in calculating large $\bar{F}_{BJ}(q)$ at small q, but also takes the advantages of fast computation of LS method under high dimension.

[Figure 1 about here.]

[Figure 2 about here.]

[Table 3 about here.]

[Figure 3 about here.]

4. Library Construction and Simulation

5. Application

6. Discussion

Put your final comments here.

Acknowledgements

The authors thank Professor A. Sen for some helpful suggestions, Dr C. R. Rangarajan for a critical reading of the original version of the paper, and an anonymous referee for very useful comments that improved the presentation of the paper.

Supplementary Materials

[Figure 4 about here.]

[Figure 5 about here.]

References

Barnett, I. J. and Lin, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika* **101,** 964–970.

Berk, R. and Jones, D. (1979). Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Probability Theory and Related Fields* **47,** 47–59.

Cai, T. T. and Wu, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory* **60,** 2217–2232.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32,** 962–994.

Eichler, E., Flint, J., Gibson, G., Kong, A., Leal, S., Moore, J. H., and Nadeau, J. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11,** 446–450.

Fisher, R. (1934). *Statistical Methods for Research Workers.* Biological monographs and manuals. Oliver and Boyd.

Huo, Z., Tang, S., Park, Y., and Tseng, G. (2019). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics* **36,** 524–532.

Jaeschke, D. (1979). The Asymptotic Distribution of the Supremum of the Standardized Empirical Distribution Function on Subintervals. *The Annals of Statistics* **7,** 108 – 115.

Khmaladze, E. and Shinjikashvili, E. (2001). Calculation of noncrossing probabilities for poisson processes and its corollaries. *Advances in Applied Probability* **33,** 702–716.

Li, J. and Siegmund, D. (2015). Higher criticism: p-values and criticism. *The Annals of Statistics* **43,** 1323 – 1350.

Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115,** 393–402. PMID: 33012899.

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annual Review of Genomics and Human Genetics* **9,** 387–402. PMID: 18576944.

Moscovich, A., Nadler, B., and Spiegelman, C. (2016). On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics* **10,** 2329 – 2354.

Noe, M. (1972). The Calculation of Distributions of Two-Sided Kolmogorov-Smirnov Type Statistics. *The Annals of Mathematical Statistics* **43,** 58 – 64.

Oh, M.-S. and Berger, J. O. (1992). Adaptive importance sampling in monte carlo integration. *Journal of Statistical Computation and Simulation* **41,** 143–168.

Rubinstein, R. Y. (2002). Cross-entropy and rare events for maximal cut and partition problems. *ACM Trans. Model. Comput. Simul.* **12,** 27–53.

Song, C. and Zhang, H. (2014). Tarv: Tree-based analysis of rare variants identifying risk modifying variants in ctnna2 and cntnap2 for alcohol addiction. *Genetic Epidemiology* **38,** 552–559.

Vázquez-Abad, F. J. and Dufresne, D. (1998). Accelerated simulation for pricing asian options. In *Proceedings of the 30th Conference on Winter Simulation*, WSC '98, page 1493–1500, Washington, DC, USA. IEEE Computer Society Press.

Zhang, H., Jin, J., and Wu, Z. (2020). Distributions and power of optimal signal-detection statistics in finite case. *IEEE Transactions on Signal Processing* **68,** 1021–1033.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

Appendix

Title of appendix

Put your short appendix here. Remember, longer appendices are possible when presented as Supplementary Web Material. Please review and follow the journal policy for this material, available under Instructions for Authors at `http://www.biometrics.tibs.org`.

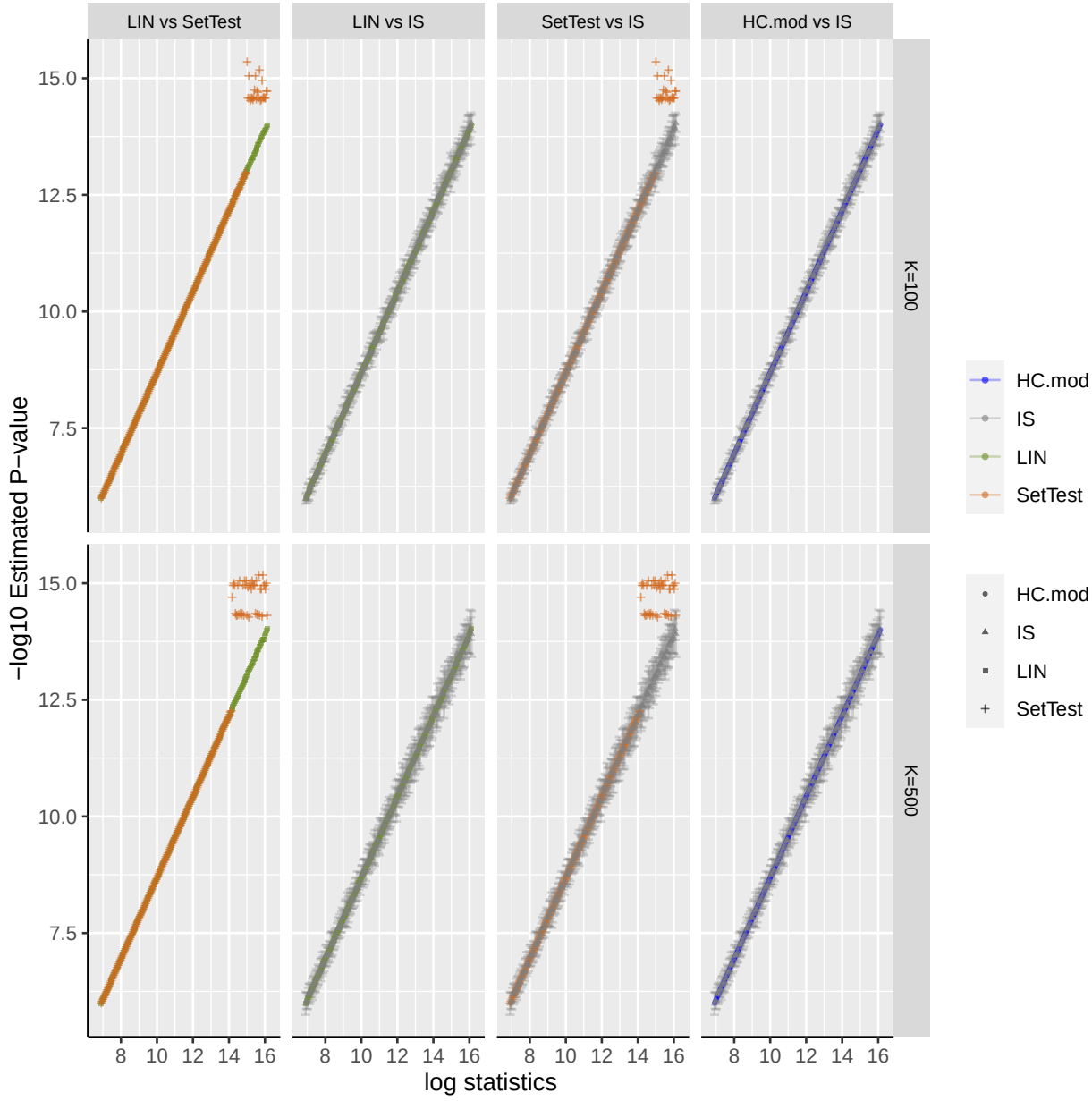

Figure 1. Methods comparison among LIN, SetTest, HC.mod and IS for FHC with significance level ranging from 10^{-14} to 10^{-6} under dimension $K = 100$ (upper) and $K = 500$ (bottom). For IS, the error bar is ±(coefficient of variation over 10 replications).

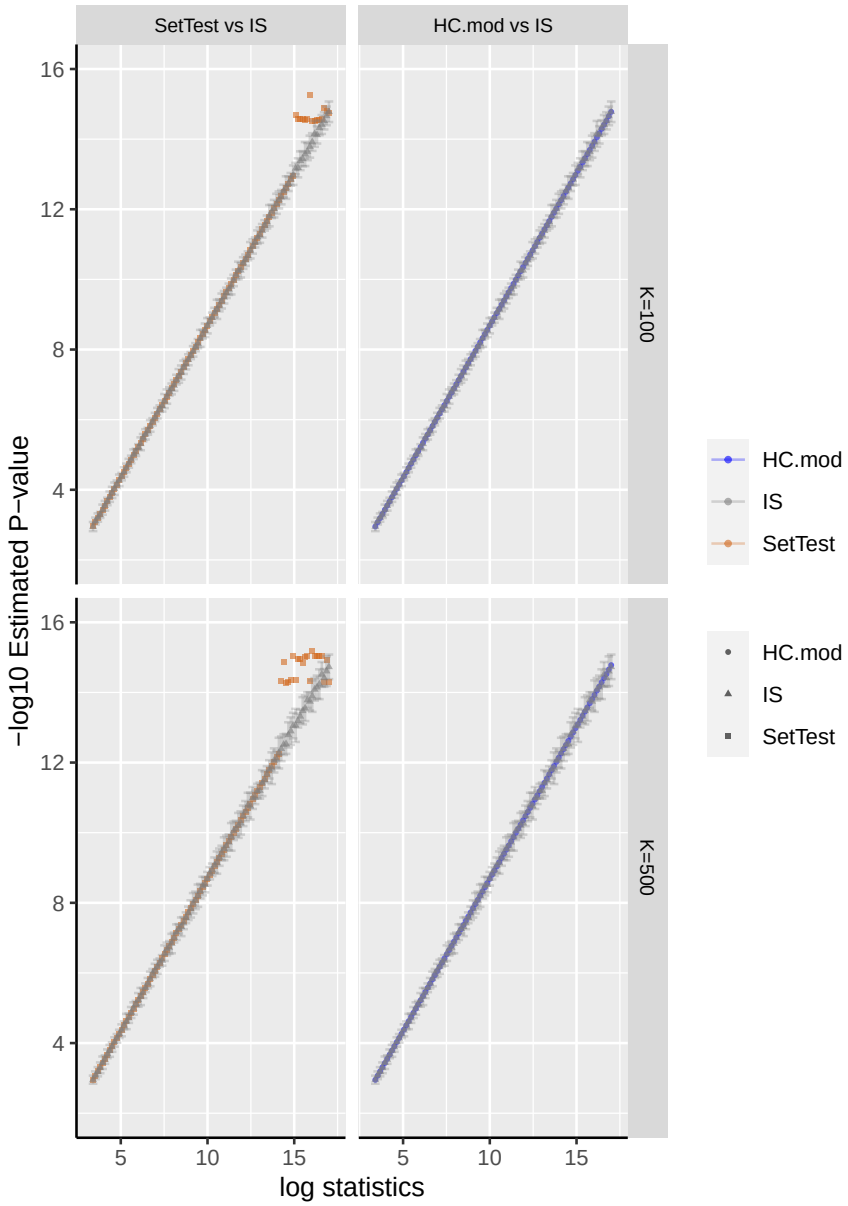

Figure 2. Methods comparison among LIN, SetTest, HC.mod and IS for HHC with significance level ranging from 10^{-15} to 10^{-3} under dimension $K = 100$ (upper) and $K = 500$ (bottom).For IS, the error bar is $\pm$(coefficient of variation over 10 replications).

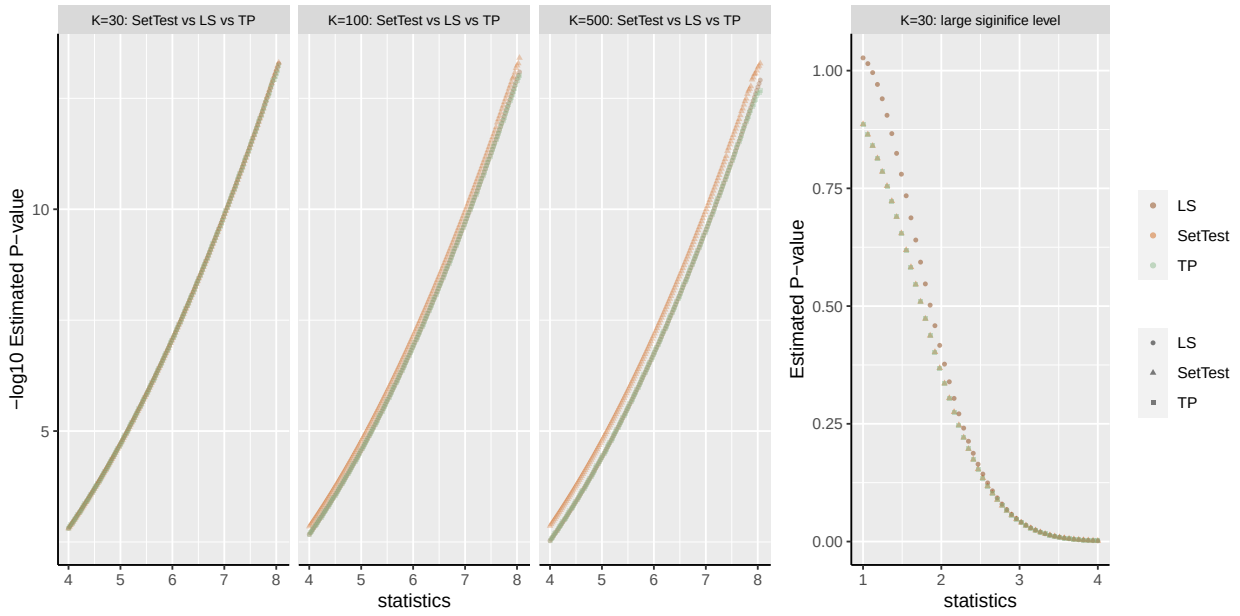

Figure 3. The left three figures compare LS, SetTest, TP for T_{BJ} at significance level ranging from 10^{-13} to 10^{-3} under dimension $K = 30$ (column 1), $K = 100$ (column 2) and $K = 500$ (column 3) respectively.The rightest plot compares their performance at significance level between 10^{-3} and 0.9.

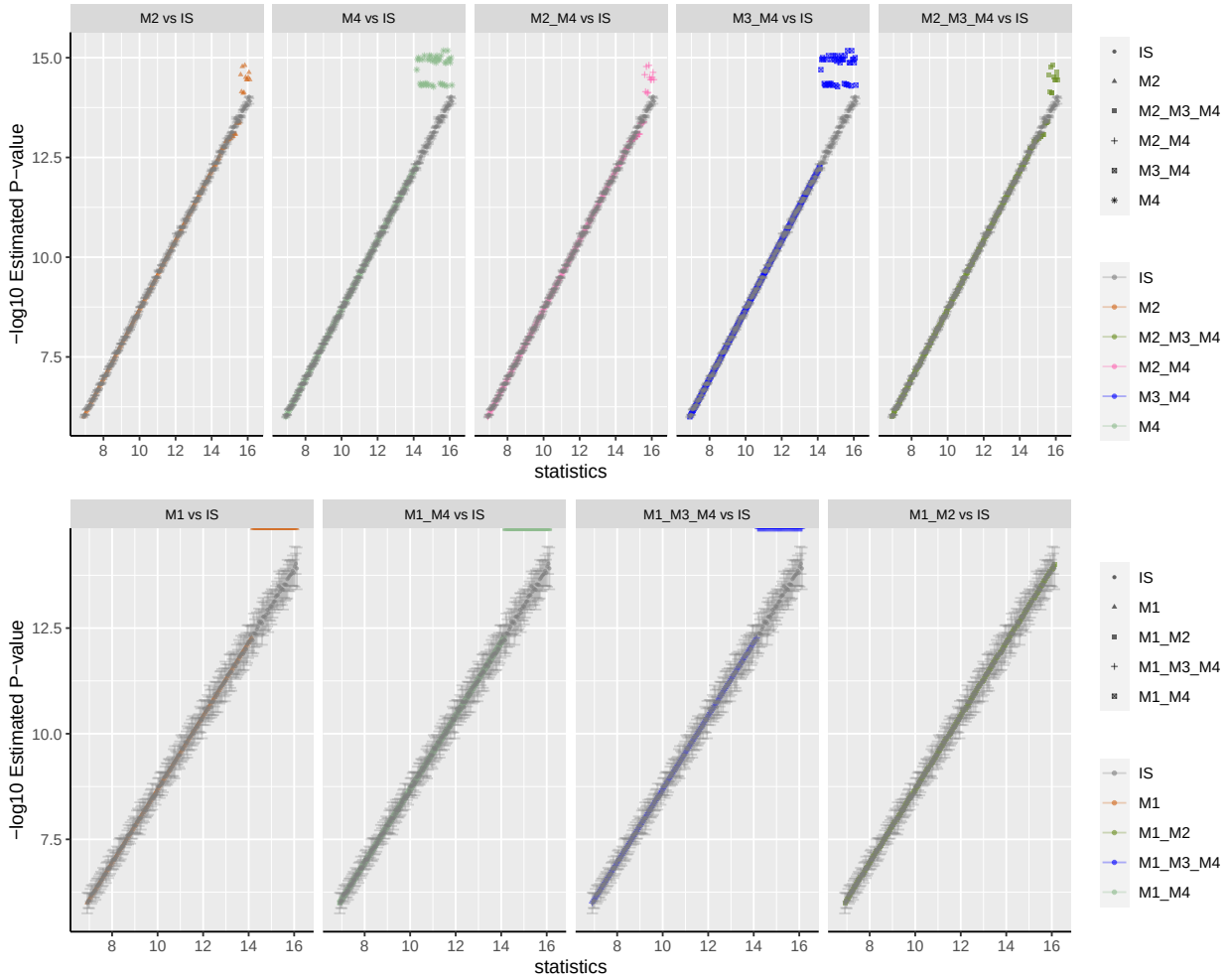

Figure 4. For FHC with $K = 500$, the top raw shows that without the Modification 1, SetTest still fails in calculating small combined p-value, which indicates the necessity of Modification 1. The bottom row demonstrates that the Modification 2 is also essential after conducting Modification 1. Furthermore, Modification 3 and Modification 4 must be conducted at the same time, otherwise the algorithm would directly break down and the justifying plots are unable to be generated. The last plot implies that Modification 1 and Modification 2 are sufficient to remedy SetTest.

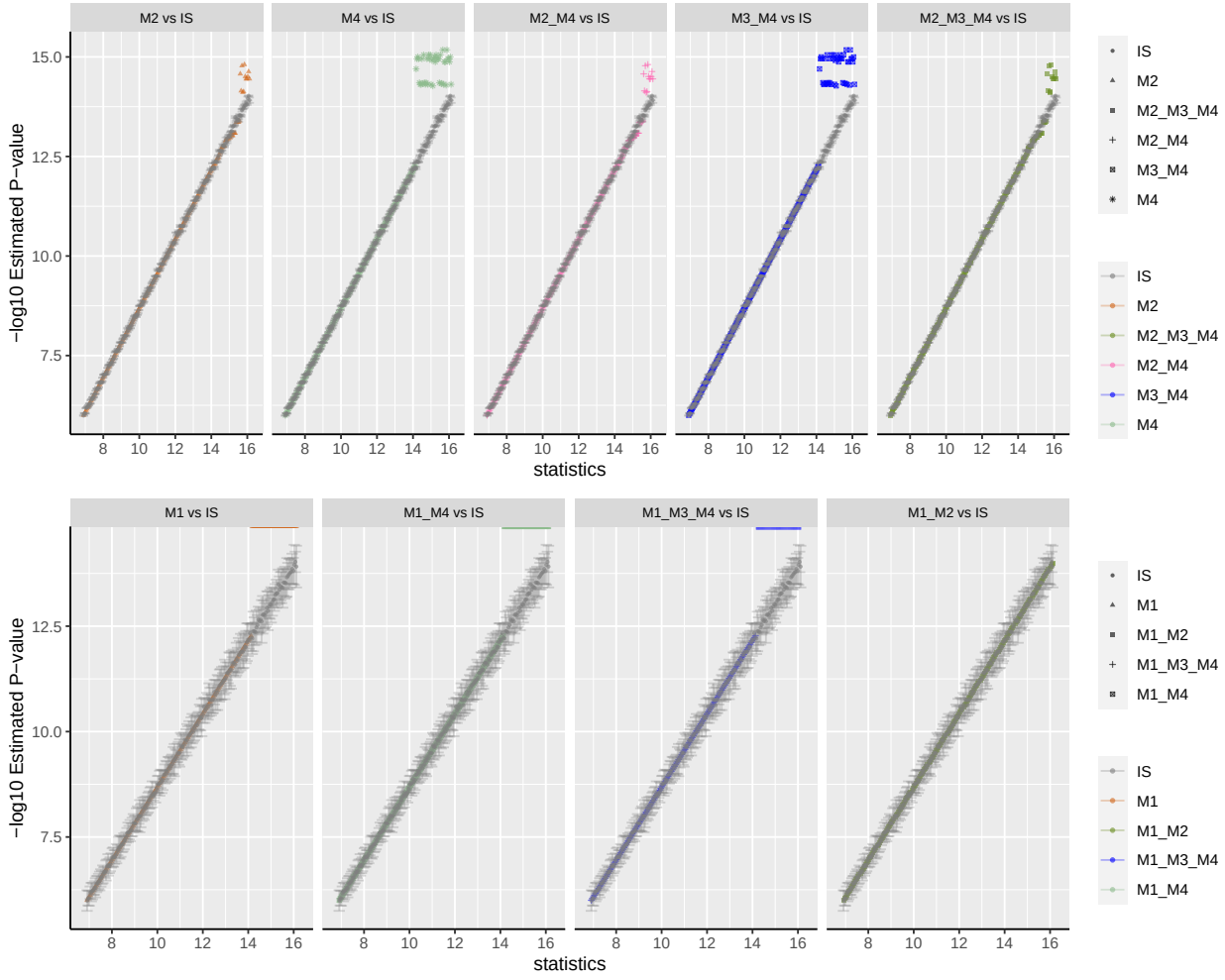

Figure 5. These are the justifying plots for HHC with $K = 500$. They show the similar patterns and have the same illustrations as Figure 4.

Class	Method	Full domain	Truncated domain	Thresholded domain
Sampling	Monte Carlo	pros: applicable for all domain. cons: unable to calculate the extremely right tail probability.		
	IS-MixGauss	pros: a. applicable for all domain. b. able to calculate the extremely right tail probability.		
Approx	Li-Siegmund	pros: fast. cons: fail to approximate the whole distribution.		inapplicable
Analytic	Barnett-Lin	pros: accurate cons: slow	inapplicable	inapplicable
	SetTest	pros: fast. cons: severe numerical errors for extremely right tail probability.		
	HC.mod	pros: fast and precise.		inapplicable

Table 1

Methods Summary for HC: SetTest is proposed by Zhang et al. (2020). HC.mod is the modified SetTest (Future exploration: theoretically HC.mod can be extended for HC with thresholded domain, but I have not explored further). IS-MixGauss is importance sampling with importance density of mixed Gaussian.

Class	Method	Full domain	Truncated domain	Thresholded domain
Sampling	Monte Carlo	pros: applicable for all domain. cons: unable to calculate the extremely right tail probability.		
	IS-MixGauss	inaccurate estimates with large variation.		
Approx	Li-Siegmund	pros: fast computation. cons: fail to approximate the whole distribution.		inapplicable
Analytic	Barnett-Lin	inapplicable	inapplicable	inapplicable
	SetTest	pros: fast. cons: inaccurate due to the numerical truncation parameter.		
	TP	pros: precise and moderately fast.	inapplicable	inapplicable

Table 2

Methods Summary for BJ: TP is the translated polynomials technique which we develop for full domain BJ to reduce numerical error acculation after removing the numerical approximation in SetTest. (Future exploration: the importance sampling method may still work by choosing a more appropriate importance density instead of mixed Gaussian)

Table 3
Comparison of time complexity.

Test		FHC			HHC			BJ	
K	50	100	500	50	100	500	50	100	500
SetTest	1.2s	1.5s	3.4s	1.4s	1.4s	2.7s	1.1s	1.2s	2.7s
HC.mod	1.9s	2.8s	21.7s	2.0s	2.3s	6.7s	*	*	*
LIN	2.6s	2.2m	22.5m	*	*	*	*	*	*
IS	2.4m	2.4m	8.6m	3.5m	2.4m	2.6m	*	*	*
TP	*	*	*	*	*	*	1.2s	1.5s	1.9m
LS	*	*	*	*	*	*	0.7s	0.6s	0.5s