# Exceedance control of the false discovery proportion via high precision inversion method of Berk Jones statistics

Jeffrey C. Miecznikowski[1,*], Jiefei Wang

**Abstract**

Exceedance control of the false discovery proportion (FDP) can provide an interpretable method for addressing the variability in the false discovery proportion estimates. Exceedance control of FDP can be viewed as constructing a confidence interval for FDP and as such inverting a hypothesis test is a viable method for achieving exceedance control. A novel powerful approach for exceedance control is presented based on using a directional Berk-Jones goodness-of-fit statistic. The approach employs a high-precision implementation procedure to accurately compute confidence envelopes for FDP. The procedure is compared against other methods and generalized to include other goodness-of-fit statistics that follow an isotropy condition.

**Keywords:** exceedance control, false discovery proportion, confidence envelope, Berk-Jones statistic

## 1. Introduction

Multiple testing procedures to control the false discovery rate (FDR) started with the BH approach presented in [1]. FDR is defined as the expected value of the false discovery proportion (FDP) which is the number of false rejections divided by the number of rejections. FDR control is often more desirable in a large-scale multiple hypothesis test setting than the family-wise error rate (FWER) control since the power with FDR is much larger than with FWER control.

While the false discovery rate controls the expectation of the false discovery proportion, in some cases, it can be useful to control the probability that the false discovery proportion exceeds a specified bound. This is what is meant by exceedance control of the false discovery proportion. Informally, exceedance control of the false discovery proportion is analogous to a confidence interval whereas FDR is analogous to a point estimate. In short, exceedance control of the false discovery proportion (often denoted as FDX) is implemented by controlling the upper limit tail probability of the false discovery proportion. The exceedance control of the false discovery proportion controls a quantile of the FDP distribution and thus offers a stronger guarantee than FDR control. Examples where exceedance control may be desirable include functional magnetic resonance imaging (MRI) experiments, mass spectromery (MS), RNA-Seq and other types of sequencing (e.g. 16S microbiome), or similar omics/high-throughput experiments.

*Corresponding Author
*Email address:* jcm38@buffalo.edu (Jeffrey C. Miecznikowski)
[1]SUNY University at Buffalo, Department of Biostatistics, 723 Kimball Tower, Buffalo, NY, 14214 USA, Tel:716-829-3446, Fax: 716-829-2200

Exceedance control of the false discovery proportion has received a good deal of attention (see [6, 15, 20, 17, 7, 10, 4, 5, 9]). There are mainly two classes of methods to control the exceedance probability, namely the augmentation approach from [20] and the inversion method from [6] and [7]. In this paper, we will focus on the inversion method, which is a post hoc method that provides the upper confidence bound of the FDP for an arbitrary set of rejections after observing the data. This property makes the method very flexible and allow researchers to adjust the hypotheses that will be rejected according to different research endpoints.

The inversion method is coined as such as it is based on inverting a hypothesis test, see, for example, Chapter 8 in [3]. Particularly, this approach of inverting a test to obtain a confidence interval thus controlling, say, the exceedance probability is presented and discussed in [6].

[6] mentions that inverting traditional uniformity tests, such as the Kolmogorov-Smirnov test do not fare well in exceedance control settings as those tests are designed to look for uniformity deviations equally through all $p$-values while reasonable procedures should focus on the left tail. In order to restrict focus on the left tail, [6] propose a method to control exceedance probability utilizing, say, the smallest 10 p-values. This approach is generalized as combining $p_{(k)}$ tests. We note this approach is reasonable however it requires the user to choose $k$, that is, the number of p-values to use in creating the confidence interval.

In this article, we present an exceedance control approach based on inverting a test derived from a directional Berk-Jones statistic. This method has improved power versus combining $p_{(k)}$ tests, is faster, and does not require the user to specify which $p$-values to combine. Thus, we feel this contribution will be useful to the statistical community.

Our manuscript is presented as follows: First we present a background section summarizing exceedance control via inversion method and the (directional) Berk-Jones statistic which can be tuned to detect specific violations of uniformity. In Section 3 we present a fast algorithm for controlling the FDR exceedance probability based on the directional Berk-Jones statistic. In Sections 4 and 5 we present Simulations and an Example, respectively. Lastly we note the availability of our `R` package in addition to Discussion and Conclusion. An Appendix contains theorems and proofs used to support our method.

## 2. Exceedance control

### *2.1. Test framework*

Consider $m$ null hypotheses $H_{01}, H_{02}, ..., H_{0m}$ and the corresponding p-values $P_1, P_2, ..., P_m$, Let $\Omega$ be the set $\{1, 2, ..., m\}$ and $T \subseteq \Omega$ be the indices of the true null hypotheses where $i \in T$ implies the null hypothesis $H_{0i}$ is true. Given the rejection index set $R$ for which the corresponding $H_{0i}$'s are rejected, the false discovery proportion (FDP) is defined by

$$\text{FDP} = \Gamma(R) = \frac{\mid R \cap T \mid}{\mid R \mid}, \tag{1}$$

where $|X|$ is the cardinality of the set $X$. Instead of controlling the false discovery rate (FDR), which is defined by the expectation of FDP, we control the exceedance rate of the false discovery proportion (FDX), that is

$$\text{FDX} = \text{pr}(\text{FDP} > c) = \text{pr}(\Gamma(R) > c), \tag{2}$$

for a given $c$. We say the exceedance rate is controlled at a level $\alpha$ if and only if

$$\text{pr}(\Gamma(R) > c) < \alpha, \tag{3}$$

for a given $\alpha$. Controlling exceedance rate bounds the probability of observing a large FDP which can offer a strong constraint on the tail behavior of FDP compared with FDR. Note that [6] and [7] have independently developed the same method for controlling the quantity in (2). They both build an upper bound for the FDP but use different terminologies to refer it. [6] call it the inversion-based confidence envelope while [7] describe it as the closed-testing-based confidence set. In this paper, we will mainly follow the framework in [6] and call the upper bound the confidence envelope.

*2.2. Confidence envelope*

An $100(1-\alpha)\%$ confidence envelope $\hat{\Gamma}_\alpha(R)$ of the false discovery proportion is defined as

$$\text{pr}\left(\Gamma(R) \leq \hat{\Gamma}_\alpha(R) \text{ for all } R \subseteq \Omega\right) \geq 1-\alpha. \tag{4}$$

The confidence envelope serves as an upper bound of FDP and simultaneously works for all rejection index sets $R$. Therefore, after observing the data, researchers are free to choose which hypotheses are rejected and obtain an estimation of the upper bound of the FDP. Given the rejection index set $R_0$, the confidence envelope $\hat{\Gamma}_\alpha(R_0)$ ensures that

$$\text{pr}\left(\Gamma(R_0) \leq \hat{\Gamma}_\alpha(R_0)\right) \geq 1-\alpha, \tag{5}$$

or, in terms of the tail probability

$$\text{pr}\left(\Gamma(R_0) > \hat{\Gamma}_\alpha(R_0)\right). < \alpha, \tag{6}$$

In practice, there are two ways to apply (5) in a multiple hypothesis test setting. One can prespecify the exceedance probability $\alpha$ and estimate the confidence envelope for the FDP for a rejection set $R_0$. Alternatively, if the error rate is given, say FDP$> c$ is undesirable, we can control the probability that the FDP exceeds $c$ by finding the largest rejection set $R$ for which $\hat{\Gamma}_\alpha(R) \leq c$.

*2.3. Inversion-based confidence envelope*

In this section, we summarize the inversion based confidence envelope proposed in [6]. The confidence envelope is built upon a p-value function from a goodness-of-fit test statistic for the uniform distribution. Assume that $W$ is an integer sequence from $\Omega$ and let $P_W$ represent the sequence $\{P_{W_1}, P_{W_2}, ..., P_{W_{|\text{W}|}}\}$. Let $\Psi_{|\text{W}|}(P_W)$ be a $|\text{W}|$-variate p-value function of a goodness-of-fit test where the null hypothesis is that the samples in $P_W$ are all from the null distribution. That is,

$$\Psi_{|\text{W}|}(P_W) = \Psi_{|\text{W}|}(P_{W_1}, P_{W_2}, ..., P_{W_{|\text{W}|}}). \tag{7}$$

and the goodness-of-fit test rejects the null at level $\alpha$ when $\Psi_{|\text{W}|}(P_W) \leq \alpha$. Let $U_\alpha$ be a collection of all sequences $W$ which are not rejected by the goodness-of-fit test at level $\alpha$. That is,

$$U_\alpha = \{W \subset \Omega \mid \Psi_{|\text{W}|}(P_W) > \alpha\}. \tag{8}$$

The inversion based $100(1-\alpha)\%$ level confidence envelope is defined as

$$\hat{\Gamma}_\alpha(R) = \begin{cases} \max_{B \subset U_\alpha} \frac{|B \cap R|}{|R|} & \text{if } R \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Furthermore, if the p-value function $\Psi_{|\mathrm{W}|}(P_W)$ is isotropic, where switching the order of the input variables does not change the value of the function (e.g. $\Psi_2(P_1, P_2) = \Psi_2(P_2, P_1)$), we can rewrite (9) using the index of the ordered p-values where the information regarding the order of $P_W$ is discarded. The isotropic condition is commonly seen in GOF testing as the samples being tested are usually unweighted and switching the order of the samples would not change the test result (e.g. Kolmogorov–Smirnov test). Let $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ be ascending ordered p-values and define $P_{(W)}$ as the sequence $\{P_{(W_1)}, P_{(W_2)}, ..., P_{(W_{|\mathrm{W}|})}\}$ where $\Psi_{|\mathrm{W}|}(P_{(W)}) = \Psi_{|\mathrm{W}|}(P_{(W_1)}, P_{(W_2)}, ..., P_{(W_{|\mathrm{W}|})})$ is the p-value function of the goodness-of-fit test. Define

$$\bar{U}_\alpha = \{W \subset \Omega \mid \Psi_{|\mathrm{W}|}(P_{(W)}) > \alpha, W_1 < W_2 < ... < W_{|\mathrm{W}|}\}, \tag{10}$$

where $W$ is an ascending ordered sequence and $\bar{U}_\alpha$ is a collection of all sequences $W$ such that the samples $P_{(W_1)}, P_{(W_2)}, ..., P_{(W_{|\mathrm{W}|})}$ are not rejected by the goodness-of-fit test at level $\alpha$. Note that we treat a sequence as a set with an order, so the set operations such as $\subset$ can be applied to a sequence as well. The $100(1-\alpha)\%$ level confidence envelope is given as

$$\hat{\Gamma}_\alpha(\bar{R}) = \begin{cases} \max_{B \subset \bar{U}_\alpha} \frac{|B \cap \bar{R}|}{|\bar{R}|} & \text{if } \bar{R} \neq \emptyset \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $\bar{R}$ is an ascending ordered rejection index sequence.

It is straightforward to observe that (8) and (10) contain the same unrejected sets of hypotheses. To be more clear, let $\pi(i)$ be a rank function defined as

$$P_i = \pi(i)^{th} \text{ smallest P-value} = P_{(\pi(i))}. \tag{12}$$

For an arbitrary p-value subset $P_W$ and its sorted version $P_{(\pi(W))}$ where $\pi(W) = \{\pi(W_1), ..., \pi(W_{|\mathrm{W}|})\}$. If the isotropic assumption is met, we have $\Psi_{|\mathrm{W}|}(P_{(\pi(W))}) = \Psi_{|\mathrm{W}|}(P_W)$ and, therefore, the sets $U_\alpha$ and $\bar{U}_\alpha$ contain the same hypotheses, except that the former set $U_\alpha$ contains the indices of the original hypotheses while the latter set $\bar{U}_\alpha$ contains the indices of the hypotheses sorted by the p-values. Consequently, the confidence envelop yields from $\Gamma_\alpha(R)$ and $\Gamma_\alpha(\bar{R})$ are equivalent for the same set of rejections.

Since $\Gamma_\alpha(R)$ and $\Gamma_\alpha(\bar{R})$ can be used to construct the same confidence envelope when the p-value function of the goodness-of-fit test is isotropic, we will focus on the goodness-of-fit tests that have isotropic property and use the confidence envelope defined in (9). To ease the notational burden, we will use $U$ and $R$ to represent $\bar{U}$ and $\bar{R}$ respectively unless otherwise stated.

In the next section, we summarize two goodness-of-fit statistics to test for the standard uniform distribution ($U(0,1)$). We are interested in using these statistics to assess whether a subset of $p$-values in Section 2.1 follow a null $U(0,1)$ distribution. The associated tests for the statistics have isotropic p-value functions.

*2.4. kth order statistic and the combined method*

The $k$th order statistic is a test statistic presented in [6] to control the exceedance rate in (2). The $k$th order statistic is based on the $k$-th smallest p-value in the subset and we denote its $p$-value function as $\Psi_{|\mathrm{W}|}^{P_{(k)}}(P_{(W)})$. Specifically, the p-value of the $k$th order statistic under the $U(0,1)$ null hypothesis is

$$\begin{aligned}\text{p-value} &= \Psi_{|\mathrm{W}|}^{P_{(k)}}(P_{(W)}) \\ &= \Psi_{|\mathrm{W}|}^{P_{(k)}}(P_{(W_1)}, P_{(W_2)}, ..., P_{(W_{|\mathrm{W}|})}) \\ &= \mathrm{pr}\left(Beta(k, |\mathrm{W}| - k + 1) < P_{(W_k)}\right) \\ &= B_{|\mathrm{W}|,k}\left(P_{(W_k)}\right),\end{aligned} \tag{13}$$

where $B_{|\mathrm{W}|,k}(x)$ is the Beta distribution function with two shape parameters $a = k$ and $b = |\mathrm{W}| - k + 1$. Using the $k$th order statistic, we define the collection of the unrejected p-value subsets as

$$U_\alpha^{P_{(k)}} = \{W \subset \Omega \mid \Psi_{|\mathrm{W}|}^{P_{(k)}}(P_{(W_1)}, P_{(W_2)}, ..., P_{(W_{|\mathrm{W}|})}) > \alpha, W_1 < W_2 < ... < W_{|\mathrm{W}|}\}. \tag{14}$$

The $k$th order statistic based confidence envelope $\hat{\Gamma}_\alpha^{P_{(k)}}(R)$ is

$$\hat{\Gamma}_\alpha^{P_{(k)}}(R) = \begin{cases} \max_{B \subset U_\alpha^{P_{(k)}}} \frac{|B \cap R|}{|R|} & \text{if } R \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The confidence envelope in (15) seems computationally unfeasible for it requires testing all subsets of the p-values $P_{(1)}, P_{(2)}, ..., P_{(m)}$ to obtain $U_\alpha^{P_{(k)}}$ in (14), but [6] provides a fast algorithm for $\hat{\Gamma}_\alpha^{P_{(k)}}(R)$ to reduce the computational complexity. Let

$$J_k = \min\{j : P_{(j)} \geq B_{|\mathrm{W}|,j}^{-1}(\alpha)\}. \tag{16}$$

The confidence envelope $\hat{\Gamma}_\alpha^{p_k}(R)$ can be computed via

$$\hat{\Gamma}_\alpha^{P_{(k)}}(R) = \begin{cases} 1 - \frac{\left|\{k, ..., J_k\} \bigcap R\right|}{|R|} & \text{if } J_k \neq \emptyset \text{ and } R \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The fast algorithm make it possible to compute the confidence envelope on a linear complexity. Furthermore, since the power of the test depends greatly on the choice of $k$, [6] developed a, so called, combined method where the idea is to combine envelopes obtained under different $k$'s to reduce the chances of a poor envelope obtained from selecting an inappropriate $k$. Let $Q \subset \Omega$ be a set of indices and the combined confidence envelope, $\hat{\Gamma}_\alpha^{P_{(Q)}}(R)$, is defined as

$$\hat{\Gamma}_\alpha^{P_{(Q)}}(R) = \min_{k \in Q} \hat{\Gamma}_{\alpha/|Q|}^{P_{(k)}}(R). \tag{18}$$

Note when $Q = \{k\}$, the confidence envelope of the combined method $\hat{\Gamma}_\alpha^{P_{(Q)}}(R)$ degrades to the confidence envelope of the $k$th order statistic $\hat{\Gamma}_\alpha^{P_{(k)}}(R)$. Note that $\hat{\Gamma}_\alpha^{P_{(Q)}}(R)$ is always greater or

equal than $\hat{\Gamma}_{\alpha}^{P_{(k)}}(R)$ with the optimal choice of $k$ in $Q$. This is because the significant level for each set on the right of (18) is $\alpha/|Q|$ and taking the minimum over $k \in Q$ will not make $\hat{\Gamma}_{\alpha}^{P_{(Q)}}(R)$ smaller than the optimal $\hat{\Gamma}_{\alpha}^{P_{(k)}}(R)$ which is tested at level $\alpha$. However, since the combined method greatly reduces the risk of choosing an inappropriate $k$, it is more flexible than the $k$th order statistic based confidence envelope in practice. In the next section, we introduce Berk-Jones statistics which can serve as a replacement for the combined method and under independence can be more powerful than the combined method.

*2.5. Berk-Jones statistics*

The Berk-Jones statistic is a goodness-of-fit test statistic that was first introduced in [2]. It compares the ascending ordered samples with its theoretical distribution under the null and the test is declared significant if and only if at least one order statistic is too large or too small. For the samples $P_{W_1}, P_{W_2}, ..., P_{W_{|\mathrm{W}|}}$, define

$$Z_{|\mathrm{W}|,i} = \mathrm{pr}\left(Beta(i, |\mathrm{W}| - i + 1) \leq P_{(W_i)}\right) = B_{|\mathrm{W}|,i}\left(P_{(W_i)}\right), \tag{19}$$

where $B_{|\mathrm{W}|,i}$ is the Beta distribution function with two shape parameter $a = i$ and $b = |\mathrm{W}| - i + 1$. The statistic $Z_{|\mathrm{W}|,i}$ often is called the local level for it carries the significance information of the "local" $P_{(W_i)}$. For the "global" significance, the BJ statistic $M_{|\mathrm{W}|}$, and its one-sided variants $M^{-}_{|\mathrm{W}|}$ and $M^{+}_{|\mathrm{W}|}$ can be defined by

$$M^{+}_{|\mathrm{W}|} = \min_{1 \leq i \leq |\mathrm{W}|} Z_{|\mathrm{W}|,i},\ M^{-}_{|\mathrm{W}|} = \min_{1 \leq i \leq |\mathrm{W}|}(1 - Z_{|\mathrm{W}|,i}),\ M_{|\mathrm{W}|} = min\{M^{+}_{|\mathrm{W}|}, M^{-}_{|\mathrm{W}|}\}. \tag{20}$$

In contrast to the well-known Kolmogorov-Smirnov statistic, Berk-Jones statistics are very sensitive to detect the deviance in two tails of the distribution function. This tail-sensitive property makes the Berk-Jones statistic very suitable to build the confidence envelope. Since the p-values from the multiple hypotheses tend to be smaller than expected under the alternative in settings with biological signals (see, beta-uniform mixtures models as in [16]) a one-sided Berk-Jones statistic $M^{+}_{w}$ is more appropriate in testing the p-values.

Certain generalizations can be made on the Berk-Jones statistics to make it more flexible. Let $Q \subset \Omega$ be an index set, a partial one-sided Berk-Jones statistic, $M^{+}_{Q}$ can be defined as

$$M^{+}_{Q} \equiv \min_{i \in Q} Z_{|\mathrm{W}|,i}. \tag{21}$$

For a small subset $W$, it is possible that there exist $i \in Q$ such that $i > |\mathrm{W}|$. In this case, we ignore the out-of-bound indices that may be present in (21). The p-value function $\Psi^{M^{+}_{Q}}_{|\mathrm{W}|}$ for the observed

samples $p_{(W)}$ and Berk-Jones statistic $m_Q^+$ is

$$\begin{aligned}
\Psi_{|\mathrm{W}|}^{M_Q^+}(p_{(W)}) &= \mathrm{pr}(M_Q^+ < m_Q^+) && (22)\\
&= \mathrm{pr}(\min_{i\in Q} Z_{|\mathrm{W}|,i} < m_Q^+) && (23)\\
&= \mathrm{pr}(\bigcup_{i\in Q} B_{|\mathrm{W}|,i}(P_{(W_i)}) < m_Q^+) && (24)\\
&= 1 - \mathrm{pr}(\bigcap_{i\in Q} B_{|\mathrm{W}|,i}(P_{(W_i)}) \geq m_Q^+) && (25)\\
&= 1 - \mathrm{pr}(\bigcap_{i\in Q} P_{(W_i)} \geq B_{|\mathrm{W}|,i}^{-1}(m_Q^+)), && (26)
\end{aligned}$$

where the exact probability in (26) depends on the joint probability of the ordered uniform statistics. Both numerical and high-precision solutions exist for solving the probability, see, for example, Chapter 9 in [19]. In this paper, we will simply assume the joint probability is available. If readers are interested in this topic, please refer to the integration method in [13], the Poisson process in [12] or Steck's determinant in [11] with details of that computation in [21].

Given $\Psi_{|\mathrm{W}|}^{M_Q^+}(p_{(W)})$, a partial one-sided Berk-Jones based confidence envelope, say $\hat{\Gamma}_\alpha^{M_Q^+}(R)$, can be computed via

$$\hat{\Gamma}_\alpha^{M_Q^+}(R) = \begin{cases} \max\limits_{B\subset U_\alpha^{M_Q^+}} \frac{|B\cap R|}{|R|} & \text{if } R \neq \emptyset \\ \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

where

$$U_\alpha^{M_Q^+} = \{W \subset \Omega \mid \Psi_{|\mathrm{W}|}^{M_Q^+}(P_{(W_1)}, P_{(W_2)}, ..., P_{(W_{|\mathrm{W}|})}) > \alpha, W_1 < W_2 < ... < W_{|\mathrm{W}|}\}. \tag{28}$$

The Berk-Jones based confidence envelope $\hat{\Gamma}_\alpha^{M_Q^+}(R)$ in (27) is very similar to the combined method $\hat{\Gamma}_\alpha^{P_{(Q)}}(R)$ in (18). By comparing (13) and (19), it can be seen that the p-value function of the $k$th order statistic is equivalent to the local level $Z_{|\mathrm{W}|,k}$ for the same subset $W$. That is, $\Psi_{|\mathrm{W}|}^{P_{(k)}}(P_{(W)}) = Z_{|\mathrm{W}|,k}$. These quantities form the starting point for building both confidence sets. The Berk-Jones based confidence envelope directly uses the minimum of the local level $Z_{|\mathrm{W}|,k}$'s as a test statistic to build the confidence envelope, while the combined method first computes the confidence envelope for each $k$th order statistic then obtains the combined result by computing the minimum of all confidence envelopes of the $k$th order statistics. It can be shown that the Berk-Jones based confidence envelope is more powerful than the combined method under independent observations (see Appendix **??**). The intuition for the improvement in power is the following: the Berk-Jones based confidence envelope directly uses the most informative (i.e. most significant) test statistic in constructing the confidence envelope. Meanwhile the combined method in [6] proposes using a range of likely informative statistics and creates Bonferroni adjusted confidence envelopes for each value in the range ultimately taking the minimum of those confidence envelopes as the final envelope. If the indices set $Q$ only contains a single value, both Berk-Jones based and combined

test based confidence envelopes yield the same results and thus the same power. As the range in the combined method gets larger thus likely including less informative statistics, the discrepancy in power between the two methods increases. Analogously, a Bonferroni adjustment for assessing significance gets less powerful (more conservative) as the number of tests increases. Similarly, the power of the combined method in constructing the confidence envelope decreases as the range of test statistics employed in construction increases. Thus, it is best to employ the combined method over a range of statistics likely to be informative, that is, likely to be true alternatives (see [6] for more details on the range selection.)

Similar to the problem with the $k$th order statistic approach, the Berk-Jones based confidence envelope is inapplicable in practice as (27) has $O(2^m)$ computational complexity. The time consumption is unacceptable for any sample size greater than 20. Therefore, a fast algorithm is required for the Berk-Jones based confidence envelope to reduce the computational load. In the next section, we will propose a general fast algorithm to compute the confidence envelope for any BJ-like statistic. The algorithm has $O(m^2)$ computational complexity and the time cost is reasonable for sample sizes up to 10,000.

## 3. Fast algorithm for computing the confidence envelope

In this section, we propose a fast algorithm for computing the confidence envelope not only for Berk-Jones statistics, but for a wide range of statistics. This includes the $k$th order statistic, the Kolmogorov-Smirnov statistic, and the higher criticism statistic. They all share a common property and we call them BJ-like statistics.

**Definition 3.1.** Given the ascending ordered samples $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$, if the acceptance region of a goodness-of-fit test statistic can be expressed as

$$l_i^{\alpha,m} \leq P_{(i)} \leq h_i^{\alpha,m} \text{ for all } i = 1, 2, ..., m, \tag{29}$$

where $\{l^{\alpha,m}\}$ and $\{h^{\alpha,m}\}$ are two non-decreasing sequences then we call that test statistic a BJ-like statistic.

It can be shown that the Berk-Jones statistics and its variants satisfy Definition 3.1. The proof including the specific sequences for $\{l^{\alpha,m}\}$ and $\{h^{\alpha,m}\}$ can be found in Appendix **??**.

Assume that $\Psi_{|\mathrm{W}|}(P_{(W)})$ is a p-value function for a BJ-like test statistic where $W$ is an ascending integer sequence from $\Omega = \{1, 2, ..., m\}$. To compute (27) in Section 2.3, we must test all subsets of $P_{(1)}, P_{(2)}, ..., P_{(m)}$ to obtain the unrejected index set $U_\alpha$ and then maximize the false discovery proportion to obtain the confidence envelope $\hat{\Gamma}_\alpha(R)$. That is,

$$\hat{\Gamma}_\alpha(R) = \max_{B \subset U_\alpha} \frac{|B \cap R|}{|R|}, \tag{30}$$

where

$$U_\alpha = \{W \subset \Omega \mid \Psi_{|\mathrm{W}|}(P_{(W)}) > \alpha, W_1 < W_2, ..., < W_{|\mathrm{W}|}\}. \tag{31}$$

Instead of maximizing $\frac{|B \cap R|}{|R|}$ directly over all possible elements in $U_\alpha$ with different sample sizes, the fast algorithm first restricts the sample size to $n$ and computes the local maximization of $\frac{|B \cap R|}{|R|}$ over the elements in $U_\alpha$ with the same sample size $n$. The confidence envelope $\hat{\Gamma}_\alpha(R)$ is then obtained

by doing a second maximization over all local maximums for $n$ in $1, 2, ..., m$. To be more specific, define $S_n$ as a collection of $W$ such that the cardinality of $W$ is $n$, that is,

$$S_n = \{W \subset \Omega \mid W_1 < W_2, ..., < W_{|\mathrm{W}|}, |\mathrm{W}| = n\}. \tag{32}$$

Let $U_{\alpha,n}$ be the set which cannot be rejected at level $\alpha$

$$\begin{aligned} U_{\alpha,n} &= \{W \in S_n \mid \Psi_n(P_{(W)}) > \alpha\} \\ &= \{W \in S_n \mid l_i^{\alpha,n} \le P_{(W_i)} \le h_i^{\alpha,n} \text{ for } \forall i = 1, 2, ..., n\}. \end{aligned} \tag{33}$$

where the last line derives from (29). Then $U_\alpha$ is a union of $U_{\alpha,n}$s where

$$\begin{aligned} U_\alpha &= \{W \subset \Omega \mid \Psi_{|\mathrm{W}|}(P_W) > \alpha, W_1 < W_2, ..., < W_{|\mathrm{W}|}\} && (34) \\ &= \bigcup_{n=1,...,m} \{W \subset \Omega \mid \Psi_{|\mathrm{W}|}(P_W) > \alpha, W_1 < W_2, ..., < W_{|\mathrm{W}|}, |\mathrm{W}| = n\} && (35) \\ &= \bigcup_{n=1,...,m} \{W \in S_n \mid \Psi_n(P_{(W)}) > \alpha\} && (36) \\ &= \bigcup_{n=1,...,m} U_{\alpha,n}. && (37) \end{aligned}$$

Therefore, given the indices of the rejected hypotheses $R$ corresponding to the ascending ordered p-values, the $100(1-\alpha)\%$ level confidence envelope can be estimated by

$$\begin{aligned} \hat{\Gamma}_\alpha(R) &= \max_{B \subset U_\alpha} \frac{|B \cap R|}{|R|} && (38) \\ &= \max_{B \subset \bigcup_{n=1,...,m} U_{\alpha,n}} \frac{|B \cap R|}{|R|} && (39) \\ &= \max_{n=1,...,m} \max_{B \subset U_{\alpha,n}} \frac{|B \cap R|}{|R|}. && (40) \end{aligned}$$

Since the set $R$ is fixed during the maximization, we will focus on the algorithm of computing the maximization of $|B \cap R|$ over $B$ in $U_{\alpha,n}$. Define two sequence $\{p^{\alpha,n}\} = \{p_1^{\alpha,n}, p_2^{\alpha,n}, ..., p_n^{\alpha,n}\}$ and $\{q^{\alpha,n}\} = \{q_1^{\alpha,n}, q_2^{\alpha,n}, ..., q_n^{\alpha,n}\}$ as

$$p_i^{\alpha,n} = \inf\{j \mid P_{(j)} \ge l_i^{\alpha,n}, j = 1, 2, ..., m\}, \tag{41}$$

$$q_i^{\alpha,n} = \sup\{j \mid P_{(j)} \le h_i^{\alpha,n}, j = 1, 2, ..., m\}. \tag{42}$$

It is shown in Appendix **??** that the below two sets are equivalent for any realization of the samples $P_1, P_2, ..., P_m$.

$$\begin{aligned} U_{\alpha,n} &= \{W \in S_n \mid l_i^{\alpha,n} \le P_{(W_i)} \le h_i^{\alpha,n} \text{ for } \forall i = 1, 2, ..., n\} \\ &\equiv \{W \in S_n \mid p_i^{\alpha,n} \le W_i \le q_i^{\alpha,n} \text{ for } \forall i = 1, 2, ..., n\}. \end{aligned} \tag{43}$$

Note that when there exists at least one empty $p_i^{\alpha,n}$ or $q_i^{\alpha,n}$, we let the set $U_{\alpha,n}$ be empty. Therefore, finding the set $U_{\alpha,n}$ is equivalent to finding all $W \in S_n$ which satisfy $p_1^{\alpha,n} \le W_1 \le q_1^{\alpha,n}$, $p_2^{\alpha,n} \le$

$W_2 \leq q_2^{\alpha,n}, \ldots, p_n^{\alpha,n} \leq W_n \leq q_n^{\alpha,n}$. However, not all $p_i^{\alpha,n}$ and $q_i^{\alpha,n}$ are attainable. Consider the following two examples

$$n = 2,\, p_1^{\alpha,n} = p_2^{\alpha,n} = q_1^{\alpha,n} = q_2^{\alpha,n} = 1, \tag{44}$$

and

$$n = 2,\, p_1^{\alpha,n} = p_2^{\alpha,n} = 1,\, q_1^{\alpha,n} = q_2^{\alpha,n} = 2, \tag{45}$$

where in the former case the set $U_{\alpha,n}$ is empty and in the latter one there is no $W \in U_{\alpha,n}$ such that $W_1 = q_1^{\alpha,n}$ or $W_2 = p_2^{\alpha,n}$. Therefore, further refinement is required to remove the unreachable $p_i^{\alpha,n}$ and $q_i^{\alpha,n}$. This can be done by a sequential selection process. Define the refined sequence $\{\overline{p^{\alpha,n}}\} = \{\overline{p_1^{\alpha,n}}, \overline{p_2^{\alpha,n}}, ..., \overline{p_n^{\alpha,n}}\}$ and $\{\overline{q^{\alpha,n}}\} = \{\overline{q_1^{\alpha,n}}, \overline{q_2^{\alpha,n}}, ..., \overline{q_n^{\alpha,n}}\}$ as

$$\begin{aligned}
\overline{p_1^{\alpha,n}} &= p_1^{\alpha,n} \\
\overline{p_2^{\alpha,n}} &= \max(\overline{p_1^{\alpha,n}} + 1, p_2^{\alpha,n}) \\
\overline{p_3^{\alpha,n}} &= \max(\overline{p_2^{\alpha,n}} + 1, p_3^{\alpha,n}) \\
&\ldots \\
\overline{p_n^{\alpha,n}} &= \max(\overline{p_{n-1}^{\alpha,n}} + 1, p_n^{\alpha,n}).
\end{aligned} \tag{46}$$

It is clear to see that at each step $\overline{q_i^{\alpha,n}}$ will select the smallest index possible for $W_i$. Similarly

$$\begin{aligned}
\overline{q_n^{\alpha,n}} &= q_n^{\alpha,n} \\
\overline{q_{n-1}^{\alpha,n}} &= \min(\overline{q_n^{\alpha,n}} - 1, q_{n-1}^{\alpha,n}) \\
\overline{q_{n-2}^{\alpha,n}} &= \min(\overline{q_{n-1}^{\alpha,n}} - 1, q_{n-2}^{\alpha,n}) \\
&\ldots \\
\overline{q_1^{\alpha,n}} &= \min(\overline{q_2^{\alpha,n}} - 1, q_1^{\alpha,n}).
\end{aligned} \tag{47}$$

It can be shown that

$$\begin{aligned}
U_{\alpha,n} &= \{W \subset S_n \mid p_i^{\alpha,n} \leq W_i \leq q_i^{\alpha,n} \text{ for } \forall i = 1, 2, ..., n\} \\
&= \{W \subset S_n \mid \overline{p_i^{\alpha,n}} \leq W_i \leq \overline{q_i^{\alpha,n}} \text{ for } \forall i = 1, 2, ..., n\},
\end{aligned} \tag{48}$$

and $\overline{p_i^{\alpha,n}}$ and $\overline{q_i^{\alpha,n}}$ are attainable given that $\overline{p_i^{\alpha,n}} \leq \overline{q_i^{\alpha,n}}$ for all $i = 1, 2, ..., n$. The proof of (48) can be found in the Appendix **??**. For the previous two examples in (3.43) and (3.44), the refined bounds are

$$\begin{aligned}
\overline{p_1^{\alpha,n}} &= 1,\, \overline{q_1^{\alpha,n}} = 0 \\
\overline{p_2^{\alpha,n}} &= 2,\, \overline{q_2^{\alpha,n}} = 1
\end{aligned} \tag{49}$$

and

$$\begin{aligned}
\overline{p_1^{\alpha,n}} &= 1,\, \overline{q_1^{\alpha,n}} = 1 \\
\overline{p_2^{\alpha,n}} &= 2,\, \overline{q_2^{\alpha,n}} = 2,
\end{aligned} \tag{50}$$

respectively. It is trivial to see that $U_{\alpha,n}$ is empty for the first example and $W = \{1, 2\}$ is the only

element in $U_{\alpha,n}$ for the second one.

Computing $\hat{\Gamma}_\alpha(R)$ can be done by first calculating the value of $\max_{B\subset U_{\alpha,n}} |B\cap R|$, then maximize $\max_{B\subset U_{\alpha,n}} |B\cap R|$ over all possible $n$. Given the subset size $n$ and rejection index $R$, we sequentially select the values of $B_n^* = \{b_1^*, b_2^*, ..., b_n^*\} \in U_{\alpha,n}$ such that

$$b_i^* = \begin{cases} \min\{r \in R \mid \max(\overline{p_i^{\alpha,n}}, b_{i-1}^* + 1) \le r \le \overline{q_i^{\alpha,n}}\} & , \{r \in R \mid \max(\overline{p_i^{\alpha,n}}, b_{i-1}^* + 1) \le r \le \overline{q_i^{\alpha,n}}\} \neq \emptyset \\ \max(\overline{p_i^{\alpha,n}}, b_{i-1}^* + 1) & , \text{otherwise,} \end{cases} \tag{51}$$

where we let $b_0^* = 0$. It can be shown that $|B_n^* \cap R| = \max_{B\subset U_{\alpha,n}} |B \cap R|$. The proof can be found in Appendix **??**. Now we propose the algorithm for computing $\hat{\Gamma}_\alpha(R)$ in Algorithm 1.

**Algorithm 1** computing $\hat{\Gamma}_\alpha(R)$

Define $FP = 0$
**for** integer $n$ from 1 to $m$ **do**
  Compute the sequence $\{l^{\alpha,n}\}$ and $\{h^{\alpha,n}\}$ for the BJ-like statistic
  Compute the sequence $\{p^{\alpha,n}\}$ and $\{q^{\alpha,n}\}$ using (41) and (42)
  If $\exists\ p_i^{\alpha,n} = \emptyset$ or $q_i^{\alpha,n} = \emptyset$ for $i \in \{1, \ldots, n\}$ then go to the next iteration
  Compute the sequence $\{\overline{p^{\alpha,n}}\}$ and $\{\overline{q^{\alpha,n}}\}$ using (46) and (47)
  If $\exists\ \overline{q_i^{\alpha,n}} < \overline{p_i^{\alpha,n}}$ then go to the next iteration
  Selecting $B_n^*$ according to (51)
  Let $FP = \max\{FP, |B_n^* \cap R|\}$
**end for**
Return $FP/|R|$ as the result of $\hat{\Gamma}_\alpha(R)$

As shown in Algorithm 1 each line inside the loop is a single-pass algorithm except the first line whose complexity depends on the statistic. Therefore, the algorithm itself has the complexity of $O(m^2)$ in total. It greatly reduces the computational requirement compared with the original algorithm in (9) and thus should be a better replacement in practice.

## 4. Simulation study

In this section we show the exceedance rate and the power of the inversion based confidence envelope with BJ statistic. The inversion based confidence envelope provides a method to estimate the upper bound of the false discovery proportion rather than directly determine the significant hypotheses. Therefore, we turn the inversion based confidence envelope into a testing procedure simply by rejecting the p-values from smallest to the largest stopping when the upper bound of the false discovery proportion exceeds the desired rate. We call the procedure with BJ statistic as BJ-based procedure.

We define the event of exceedance as the false discovery proportion exceeding 0.1 and set the desired exceedance rate to 0.1, that is, we want to control

$$\text{pr}(\text{FDP} > 0.1) < 0.1. \tag{52}$$

This exceedance control will be used across the entire simulations unless otherwise mentioned. The test statistics are simulated from a $m$-variate normal distribution

$$\begin{bmatrix} X_1 \\ X_2 \\ ... \\ X_m \end{bmatrix} \sim \mathrm{N}\left(\mu, \Sigma\right), \tag{53}$$

where $\mu_i = 0$ if $X_i$ is from the null and $\mu_i = \theta$ if $X_i$ is from the alternative. For simplicity, the variance for the individual $X_i$ is assumed to be a constant 1. The p-values are obtained by the right tail probability of the standard normal distribution at the observed test statistic. For the dependent data, we use two types of the covariance matrix, namely compound symmetric (CS) and an autoregressive process of order 1 (AR1). The former one simulates the case which the test statistics have consistently large correlation and the later one mimics, say, a genomics study where the correlation between two tests (genes) decreases as the distance between the genes increases. For any pairs of $X_i$ and $X_j$, the CS has the covariance

$$\mathrm{cov}(X_i, X_j) = \begin{cases} 1, & i = j \\ \rho, & i \neq j, \end{cases} \tag{54}$$

and AR1 is

$$\mathrm{cov}(X_i, X_j) = \rho^{|i-j|}, \tag{55}$$

where $\rho$ a parameter controlling the strength of the correlation. For example, if $m = 4$, the covariance matrix of CS is

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \tag{56}$$

and AR1 is

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}. \tag{57}$$

In this setting, the p-values $P_1, P_2, \ldots, P_m$ are obtained by computing the right tail probability for the sampled values under the standard normal distribution. It follows that the marginal distribution of p-value $P_i$ is given by

$$G = (1 - a)U + aF,$$

where $U(p) = p$ and $F(p)$ is a distribution obtained by computing the probability under a standard normal for random variables obtained from a normal distribution centered at $\theta$. The proportion of true alternatives is given by $a$.

Unless otherwise mentioned, we choose the effect size $\theta = 1.5$ and the proportion of true alternative $a = 0.5$ across all simulations. This represents the situation where there are weak but common signals among all tests. The testing procedures being compared include the directional Berk-Jones statistic (denoted as BJ+), the combined method with an estimated $k$ proposed in [6] denoted as CB($1 \sim \hat{k}$), and the interpolation-based method for the sorted p-values in Theorem 1 of

| | Exceedance rate | | | Power | | |
|---|---|---|---|---|---|---|
| $m$ | BJ+ all | CB $(1-\widehat{k})$ | KR | BJ+ all | CB $(1-\widehat{k})$ | KR |
| 100 | 0.01 | 0.02 | 0 | 0.015 | 0.024 | 0 |
| 500 | 0 | 0.01 | 0 | 0.02 | 0.005 | 0.014 |
| 1000 | 0 | 0 | 0 | 0.078 | 0.003 | 0.037 |
| 2000 | 0 | 0 | 0.01 | 0.195 | 0.017 | 0.083 |
| 5000 | 0 | 0 | 0 | 0.285 | 0.088 | 0.089 |

Table 1: Exceedance probability and power for each statistic for the independent data with $a = 0.5$ and $\theta = 1.5$.

| | Exceedance rate | | | Power | | |
|---|---|---|---|---|---|---|
| $\rho$ | BJ+ all | CB $(1-\widehat{k})$ | KR | BJ+ all | CB $(1-\widehat{k})$ | KR |
| 0.1 | 0.235 | 0.049 | 0.023 | 0.242 | 0.073 | 0.083 |
| 0.2 | 0.321 | 0.093 | 0.043 | 0.331 | 0.116 | 0.105 |
| 0.3 | 0.374 | 0.119 | 0.062 | 0.385 | 0.146 | 0.123 |
| 0.4 | 0.446 | 0.15 | 0.08 | 0.469 | 0.183 | 0.15 |
| 0.5 | 0.469 | 0.167 | 0.073 | 0.493 | 0.197 | 0.155 |
| 0.6 | 0.527 | 0.161 | 0.071 | 0.553 | 0.203 | 0.162 |
| 0.7 | 0.587 | 0.162 | 0.075 | 0.612 | 0.215 | 0.172 |
| 0.8 | 0.638 | 0.163 | 0.058 | 0.671 | 0.228 | 0.171 |
| 0.9 | 0.757 | 0.182 | 0.053 | 0.758 | 0.265 | 0.196 |
| 1 | 0.958 | 0 | 0.03 | 0.958 | 0 | 0 |

Table 2: Exceedance probability and power for each statistic for the CS dependent data with $m = 1000$, $a = 0.5$, and $\theta = 1.5$.

[9] (denoted as KR).

For the independent data, the covariance matrix $\Sigma$ is the identity matrix $I$. The sample sizes are $m = 100, 500, 1000, 2000, 5000$ and the simulation will be repeated 1000 times to estimate the exceedance rate and power.

Table 1 shows the the exceedance and power for each testing procedure under independence. We see all methods control the exceedance rate and the power of Berk-Jones directional statistic BJ+ outperforms the other methods except for $m = 100$.

Tables 2 and 3 show the power and exceedance rate for all methods under the CS and AR1 dependent data, respectively. We fix $m = 1000$, $a = 0.5$ and $\theta = 1.5$ and vary the correlation factor $\rho$ to see the effect of the correlation. The result shows the BJ method can tolerate a mild correlation, but fails to control the exceedance rate under a strong correlation. This is expected as the data violates the independence assumption required for the beta distribution assumption in the Section 2.5. However, interestingly, the combined method still controls the exceedance rate under 0.1 in our simulation. Since the combined method depends on a beta distribution assumption as well, this result is rather surprising. In particular, we speculate that employing a Bonferroni-type correction to create the confidence interval in the combined method provides a degree of tolerance for correlated observations. Under AR1 dependency as shown in Table 3, we see the BJ+ method controls the Type I error for small and modest values of $\rho$ however it is not as powerful as the combined method under most settings for $\rho$.

| | Exceedance rate | | | Power | | |
|---|---|---|---|---|---|---|
| $\rho$ | BJ+ all | CB $(1-\widehat{k})$ | KR | BJ+ all | CB $(1-\widehat{k})$ | KR |
| 0.1 | 0.005 | 0 | 0.003 | 0.087 | 0.005 | 0.045 |
| 0.2 | 0.004 | 0.001 | 0.001 | 0.085 | 0.005 | 0.048 |
| 0.3 | 0.006 | 0.002 | 0 | 0.084 | 0.006 | 0.045 |
| 0.4 | 0.003 | 0 | 0 | 0.096 | 0.007 | 0.051 |
| 0.5 | 0.013 | 0 | 0.003 | 0.097 | 0.009 | 0.048 |
| 0.6 | 0.021 | 0.002 | 0.004 | 0.109 | 0.012 | 0.053 |
| 0.7 | 0.044 | 0.003 | 0.008 | 0.123 | 0.018 | 0.056 |
| 0.8 | 0.081 | 0.003 | 0.008 | 0.143 | 0.027 | 0.061 |
| 0.9 | 0.193 | 0.013 | 0.022 | 0.215 | 0.046 | 0.068 |
| 1 | 0.961 | 0 | 0.012 | 0.961 | 0 | 0 |

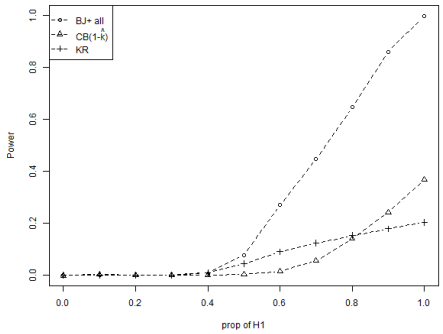
Table 3: Exceedance probability and power for each statistic for the AR1 dependent data with $m = 1000$, $a = 0.5$, and $\theta = 1.5$.

| m | CB all | BJ+ all |
|---|---|---|
| 50 | 0.065 | 0.015 |
| 100 | 0.23 | 0.07 |
| 200 | 1.01 | 0.235 |
| 500 | 6.76 | 1.475 |
| 1000 | 30.115 | 6.38 |
| 2000 | 152.6 | 31.415 |
| 5000 | 1455.405 | 306.735 |
| 10000 | 9437.645 | 1961.445 |

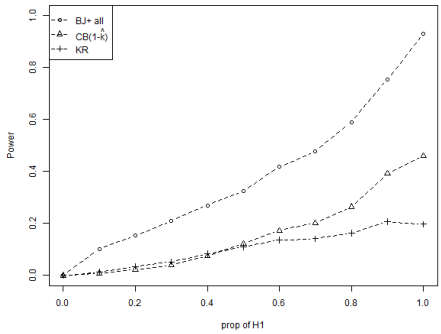
Table 4: The time consumption for the combined and Berk-Jones directional statistic measured in second

Next we vary the proportion of true alternatives $a$ to see the changes in power when the effect size and sample size are fixed at $\theta = 1.5$, $m = 1000$, respectively. The data are simulated from both independent and dependent structures with $\rho = 0.2$. Figure 1 Panels (a), (b) and (c) shows the power curve as a function of $a$ for independence, CS and AR1, respectively. As the proportion of true alternatives increases, all three procedures have an increased power. The procedure for BJ+ has the highest power among all procedures. The combined and KR procedures are comparable.

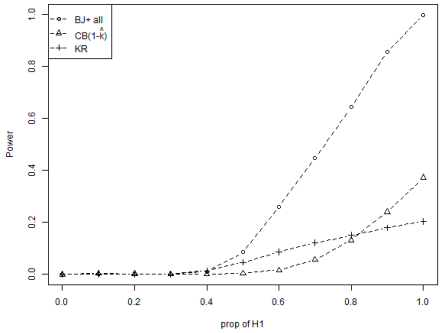
To further examine the performance of our method, we examined the computational time. We choose the sample size $m = 50, 100, 200, 500, 1000, 2000, 5000, 10000$ and use 100 iterations to compute the average time cost for each algorithm. The simulation is run on an Intel i7-8750H at 2.2GHz, 16GB memory laptop. Table 4 gives the result for the time consumption. It can be seen that both methods take a trivial time for $m$ less than 100. As the sample size increases, the demand for the computation resources increases quickly. For $m = 10000$, the combined method takes over 2 hours while our proposed method takes about half hour. The results indicate that even though the BJ+ test is slower than a single $k$th statistic test (not shown in the table), the ability to test all samples at once avoids the need to combine individual confidence levels procedure and thus makes our method faster overall compared to the combined method over the same region of p-values.




(a) Independence




(b) CS




(c) AR1

Figure 1: The power for each statistic as a function of the proportion of true alternatives $a$ with $m = 1000$ and $\theta = 1.5$.

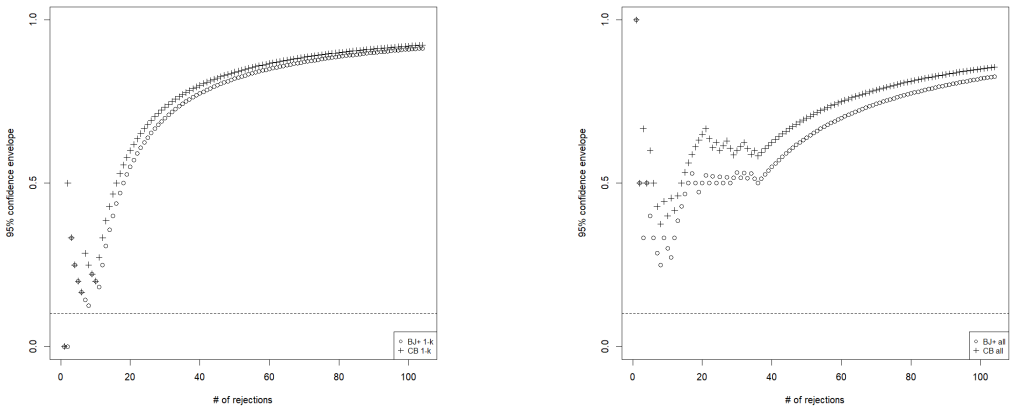
## 5. Example

We demonstrate our approach on oral cavity microbiome data designed to explore possible links between microbiome, diabetes and obesity in children. Adverse health outcomes related to diabetes and obesity are of increasing concern in the pediatric population. In adults with these diagnoses, there is a higher incidence of periodontal disease, as represented by gingivitis and periodontitis - a severe form of gum disease. For our analysis, there were 49 child subjects in this study with 19 categorized as normal weight, 14 as obese, and 16 as obese with type 2 diabetes (T2D). Each subject contributed a saliva sample for bacterial microbiome analysis via 16S rRNA sequencing. For a full analysis of these data we refer you to [8]. Here we examine 118 unique genus level genera where the genus counts are generated by aggregating the species level organizational taxonomic unit (OTU) reads under each genus. A trimmed mean of M-values (TMM) normalization is employed with tagwise dispersion parameter estimates used in a negative binomial generalized linear model to assess differential OTU bacterial abundance between control, obese, and T2D groups. For each genus level OTU the null hypothesis is no abundance difference exists between the groups and the hypothesis is evaluated via a likelihood ratio test. Figure 2 shows the BJ-based and combined-based confidence envelope that sequentially rejects null hypotheses from the most promising to the less promising p-values. Panel (a) is based on a subset region of $p$-values showing similar performance of both methods and Panel (b) is based on using all of the p-values where BJ+ has a tighter confidence envelope suggesting improved power. The tests are done at $c = \alpha = 0.1$ and the p-value region employed is $1 \sim \widehat{k}$ with $\widehat{k} = 3$ or the entire region. Using BJ+ based on a $1 \sim \widehat{k}$ region of $p$-values leads to significance for the genus level OTUs Lautropia and Scardovia, while the combined method only yields significance for Lautropia. Both of the OTUs were also determined to be significant in an FDR analysis in [8]. In assessing the direction of significance, we find Lautropia is more abundant in the control subjects relative to the T2D, while Scardovia was less abundant in controls relative to T2D subjects (boxplots shown in [8]). More recent work in oral health and microbiome presented in [14] support the significant role of Scardovia in oral health in children. Specifically, [14] determine that Scardovia is more abundant in subjects with periodontitis relative to controls which supports our results that Scardovia is more abundant in subjects with generally poorer oral health, i.e. the obese and T2D subjects. Further results in [18] also support our conclusions wherein Lautropia was significantly more abundant in a control population relative to a population with aggressive periodontitis.

## 6. Discussion and Conclusion

Exceedance control of the false discovery proportion (FDP) can provide an interpretable method for addressing the variability in the false discovery proportion estimates. Exceedance control of FDP can be viewed as constructing a confidence interval for FDP and as such inverting a hypothesis test is a viable method for achieving exceedance control. This manuscript presents a novel powerful approach for exceedance control based on using a directional Berk-Jones goodness-of-fit statistic. The method employs a fast algorithm to accurately compute our confidence envelopes for FDP. We discuss and compare our procedure against other methods and generalize our high precision approach to include other goodness-of-fit statistics that follow an isotropy condition.

Some limitations to our method are maintaining reasonable power in settings with a small number of true alternatives and a loss of error control under high levels of dependency. Also there is a high computational burden for an extremely large number of tests (e.g. for over $10,000$




(a) (b)

Figure 2: Combined-based confidence envelope made by sequentially rejecting null hypotheses from the most promising to the least promising p-values. The horizontal dashed lines indicates the 0.1 confidence envelope. The CB and BJ+ approach is based on a p-value region from $1 \sim \widehat{k} = 1 - 3$ in Panel (a) and a region using all $p$-values in Panel (b).

hypothesis tests). A further limitation includes the setting where p-values arrive in a stream with decisions to accept or reject needing to be made on the fly prior to seeing future data. We note methods presented in [9] are able to address that setting.

## 7. Code availability

All the R and C++ code for implementing the algorithms and generating the simulation results can be found at https://github.com/Jiefei-Wang/exceedance-paper.

## 8. Supplementary material

Supplementary material available online includes the proofs of the fast algorithm and sequential selection procedure.

## References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[2] Robert H Berk and Douglas H Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

[3] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[4] Sylvain Delattre, Etienne Roquain, et al. New procedures controlling the false discovery proportion via Romano–Wolf's heuristic. *The Annals of Statistics*, 43(3):1141–1177, 2015.

[5] Sebastian Döhler and Etienne Roquain. Controlling false discovery exceedance for heterogeneous tests. *arXiv preprint arXiv:1912.04607*, 2019.

[6] Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006.

[7] Jelle Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science - STAT SCI*, 26, 08 2012.

[8] Waleed F Janem, Frank A Scannapieco, Amarpeet Sabharwal, Maria Tsompana, Harvey A Berman, Elaine M Haase, Jeffrey C Miecznikowski, and Lucy D Mastrandrea. Salivary inflammatory markers and microbiome in normoglycemic lean and obese children compared to obese children with type 2 diabetes. *PLoS One*, 12(3):e0172647, 2017.

[9] Eugene Katsevich and Aaditya Ramdas. Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, 48(6):3465–3487, 2020.

[10] Erich Leo Lehmann and Joseph P Romano. Generalizations of the familywise error rate. In *Selected Works of EL Lehmann*, pages 719–735. Springer, 2012.

[11] Jeffrey C Miecznikowski, Jiefei Wang, Daniel P Gaile, and David L Tritchler. A novel exact method for significance of higher criticism via steck's determinant. *Statistics & Probability Letters*, 130:105–110, 2017.

[12] Amit Moscovich and Boaz Nadler. Fast calculation of boundary crossing probabilities for Poisson processes. *Statistics & Probability Letters*, 123:177–182, 2017.

[13] Amit Moscovich, Boaz Nadler, Clifford Spiegelman, et al. On the exact Berk-Jones statistics and their $p$-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[14] Jun Pei, Fei Li, Youhua Xie, Jing Liu, Tian Yu, and Xiping Feng. Microbial and metabolomic analysis of gingival crevicular fluid in general chronic periodontitis patients: lessons for a predictive, preventive, and personalized medical approach. *EPMA Journal*, 11(2):197–215, 2020.

[15] Marco Perone Pacifico, Christopher Genovese, Isabella Verdinelli, and Larry Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004.

[16] Stan Pounds and Stephan W Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.

[17] Sanat K Sarkar et al. Stepup procedures controlling generalized fwer and generalized fdr. *The Annals of Statistics*, 35(6):2405–2420, 2007.

[18] Meng Shi, Yiping Wei, Yong Nie, Cui Wang, Fei Sun, Wenting Jiang, Wenjie Hu, and Xiaolei Wu. Alterations and correlations in microbial community and metabolome characteristics in generalized aggressive periodontitis. *Frontiers in Microbiology*, 11:3041, 2020.

[19] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

[20] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[21] Jiefei Wang and Jeffrey C Miecznikowski. High precision implementation of Steck's recursion method for use in goodness-of-fit tests. *Journal of Applied Statistics*, 2020.