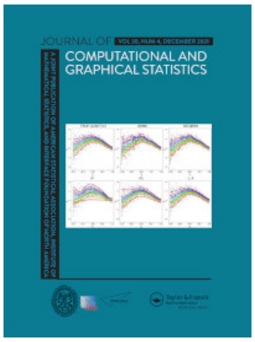

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Standardized Partial Sums and Products of *p*-Values

N. A. Heard

To cite this article: N. A. Heard (2021): Standardized Partial Sums and Products of *p*-Values, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2021.1999822

To link to this article: https://doi.org/10.1080/10618600.2021.1999822

Published online: 21 Dec 2021.

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2021.1999822

Taylor & Francis
Taylor & Francis Group

Standardized Partial Sums and Products of *p*-Values

N. A. Heard

Department of Mathematics, Imperial College London, London, UK

ABSTRACT

In meta analysis, a diverse range of methods for combining multiple *p*-values have been applied throughout the scientific literature. For sparse signals where only a small proportion of the *p*-values are truly significant, a technique called *higher criticism* has previously been shown to have asymptotic consistency and more power than Fisher's original method. However, higher criticism and other related methods can still lack power. Three new, simple to compute statistics are now proposed for detecting sparse signals, based on standardizing partial sums or products of *p*-value order statistics. The use of standardization is theoretically justified with results demonstrating asymptotic normality, and avoids the computational difficulties encountered when working with analytic forms of the distributions of the partial sums and products. In particular, the standardized partial product demonstrates more power than existing methods for both the standard Gaussian mixture model and a real data example from computer network modeling.

ARTICLE HISTORY
Received November 2019
Revised July 2021

KEYWORDS
Fisher's method; Higher criticism; Meta-analysis; Truncated product

1. Introduction

Some meta-analyses aim to combine *p*-values from multiple, approximately independent but related significance tests into an overall, global *p*-value. Let $p_1, \dots, p_n$ be *p*-values from n independent tests, with null hypothesis joint density

$$H_0 : f_n(p_1, \dots, p_n) = \prod_{i=1}^{n} \mathbb{1}_{[0,1]}(p_i). \tag{1}$$

Birnbaum (1954) showed that any statistic $T_n(p_1, \dots, p_n)$ which is monotonic must be most powerful against some alternative hypothesis, although in most cases these alternatives are difficult to relate to practical examples. The study of methods of combining *p*-values dates back to Fisher (1929), who proposed the canonical test statistic $-\sum_{i=1}^{n} \log p_i$, which conveniently follows a $\Gamma(n, 1)$ distribution under (1). Since then, several seminal approaches have appeared across the scientific literature, some with identifiable optimality properties (Heard and Rubin-Delanchy 2018).

In the modern era of routine high-dimensional data collection and high-throughput screening, it is now common for the number of tests, n, being combined in a meta-analysis to be very large, with an associated expectation that only a small proportion might be significant. Formally, this can translate to an alternative hypothesis where *p*-values are draws from a mixture density,

$$H_1 : f_n(p_1, \dots, p_n) = \prod_{i=1}^{n} \{(1 - \epsilon_n)\mathbb{1}_{[0,1]}(p_i) + \epsilon_n f_{1,n}(p_i)\}, \tag{2}$$

where the mixture proportion $0 < \epsilon_n \ll 1$; $f_{1,n}(p)$ is the density of a random variable on $[0, 1]$ which is stochastically smaller than U$[0, 1]$ and typically non-increasing in p (Birnbaum 1954). If $(\epsilon_n, f_{1,n})$ were known, by the Neyman-Pearson lemma the uniformly most powerful test would be a monotonic function of $\sum_{i=1}^{n} \log\{1 - \epsilon_n + \epsilon_n f_{1,n}(p_i)\}$; furthermore, if the indices $I_n \subseteq \{1, \dots, n\}$ of *p*-values drawn from $f_{1,n}$ were also known, an optimal statistic would be $-\sum_{i \in I_n} \log f_{1,n}(p_i)$. So, for example, if the alternative hypothesis was a mixture of uniform and Beta$(a, 1)$ *p*-values, with $a < 1$ giving a decreasing density, then Fisher's method restricted to I_n, $-\sum_{i \in I_n} \log p_i$, would be optimal.

When testing against (2), since I_n is unknown it can be convenient to construct test statistics $T_n(p_1, \dots, p_n)$ which are functions of the corresponding order statistics $p_{(1)} < \cdots < p_{(n)}$, since the lowest *p*-values are most likely to be draws from $f_{1,n}$. This motivates consideration of three primitive statistics, for $k \leq n$, which compute partial products or sums of the first k order statistics of n *p*-values:

$$s_k^n := -\sum_{i=1}^{k} \log p_{(i)}; \quad \tilde{s}_k^n := -\sum_{i=1}^{k} \log(1 - p_{(i)}); \quad \bar{s}_k^n := \sum_{i=1}^{k} p_{(i)}. \tag{3}$$

Standardizing (3) and then optimizing over k yields the following proposed test statistics:

$$\begin{aligned} \mathrm{PP}_n &= \min_{1 \leq k \leq n} \{\mathbb{E}(s_k^n) - s_k^n\}/\sqrt{\mathbb{V}(s_k^n)}, \\ \mathrm{PCP}_n &= \min_{1 \leq k \leq n} \{\tilde{s}_k^n - \mathbb{E}(\tilde{s}_k^n)\}/\sqrt{\mathbb{V}(\tilde{s}_k^n)}, \\ \mathrm{PS}_n &= \min_{1 \leq k \leq n} \{\bar{s}_k^n - \mathbb{E}(\bar{s}_k^n)\}/\sqrt{\mathbb{V}(\bar{s}_k^n)}. \end{aligned} \tag{4}$$

These correspond to standardized partial products (PP), complementary products (PCP), and sums (PS) of the k smallest *p*-values, minimized with respect to k.

CONTACT Nicholas A. Heard n.heard@imperial.ac.uk Imperial College London, Mathematics, South Kensington Campus, London SW7 2AZ, UK.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

1.1. Higher Criticism

An important, earlier contribution on combining p-values came from Donoho and Jin (2004), who developed a test statistic for sparse signals called *higher criticism*, notably derived from an earlier idea of Tukey. Their approach takes each order statistic $p_{(i)}$ in turn and, assuming (1), uses a Gaussian approximation to the binomial probability of observing i p-values not exceeding $p_{(i)}$ as a measure of the smallness of $p_{(i)}$. In its simplest form, the higher criticism statistic is

$$\mathrm{HC}_n := \max_{1\le i\le n} \mathrm{HC}_{n,i}, \quad \mathrm{HC}_{n,i} := \frac{i - np_{(i)}}{\sqrt{np_{(i)}(1-p_{(i)})}}, \tag{5}$$

which, by the monotonicity of the standard Gaussian cumulative distribution function Φ, is equivalent to finding the smallest Gaussian-approximated binomial probability. Since HC_n has no closed-form null distribution, Monte Carlo simulations are required for each n.

To motivate (5), Donoho and Jin (2004) considered the following collection of null and alternative hypotheses for n significance tests:

$$H_{0,i}: \ t_i \sim \mathrm{N}(0,1), \quad H_{1,i}: \ t_i \sim \mathrm{N}(\mu_n, 1), \tag{6}$$

yielding corresponding p-values $p_i = 1 - \Phi(t_i)$, $i = 1, \ldots, n$. The number of cases for which $H_{1,i}$ is true is assumed to be small, implying a global hypothesis test (2) with alternative p-value density $f_{1,n}(p) = \exp\{\Phi^{-1}(1-p)\mu_n - \mu_n^2\}$ on $[0,1]$. The alternative hypothesis mean was assumed to grow slowly with n, parameterized as

$$\mu_n = \sqrt{2r\log n} \tag{7}$$

for some fixed $0 < r < 1$. In contrast, to provide a sparse signal the mixture proportion ϵ_n in (2) was assumed to decrease with n, parameterized as

$$\epsilon_n = n^{-\beta} \tag{8}$$

for $0.5 < \beta < 1$, implying $\mathbb{E}(|I_n|) = n^{1-\beta} < \sqrt{n}$. This model has subsequently been referred to as the Asymptotic Rare/Weak (ARW) model (Donoho and Jin 2015).

Defining $\rho^*(\beta) = \beta - 0.5$ if $\beta \le 0.75$ and $\rho^*(\beta) = (1 - \sqrt{1-\beta})^2$ otherwise, Donoho and Jin (2004) showed that for $\rho^*(\beta) < r < 1$, as $n \to \infty$ Fisher's method cannot separate H_0 from H_1, whereas higher criticism (5) separates H_0 from H_1 with probability 1.

Despite this valuable asymptotic property, Donoho and Jin (2015) noted that $\mathrm{HC}_{n,i}$ in (5) can be poorly behaved for small i and recommend a modified version,

$$\mathrm{HC}_n^+ := \max_{1\le i\le n:\, p_{(i)} > \frac{1}{n}} \mathrm{HC}_{n,i}, \tag{9}$$

to reduce sensitivity to very small p-values. However, it will be shown through ARW model simulations and also in practice with real data, methods such as (5) or (9) which pivot on the number of p-values lying below a threshold can still lack power, as they are insensitive to small changes in even the most significant p-values.

1.2. Other Methods

Two highly cited methods acting on the smallest observed p-values will be included in later comparisons. First, Simes (1986) proposed the statistic

$$A_n := \min_{1\le i\le n} \frac{np_{(i)}}{i}, \tag{10}$$

showing with an elegant inductive proof that, like the original p-values, A_n again has uniform density on $[0,1]$ under H_0 (1). The statistic (10) can be seen to be equivalent to the so-called *false discovery rate* procedure of Benjamini and Hochberg (1995) used in multiple hypothesis testing.

Second, Zaykin et al. (2002) suggested a truncated product method (TPM); for a fixed threshold $0 < \tau \le 1$, the proposed TPM statistic was

$$W_n := -\sum_{i=1}^{n} \mathbb{1}_{[0,\tau]}(p_i) \log p_i. \tag{11}$$

A default setting of 0.05 for the threshold parameter τ was recommended, so that only p-values which are significant at the nominal 5% level are included in the summation. The TPM statistic has a convenient closed-form distribution function, derived as a mixture of a binomial distribution on the number of p-values below τ and a log-gamma distribution for the corresponding product.

More recently, Li and Siegmund (2015) proposed a p-value combination method which is attributed to an adaption of the goodness-of-fit test statistics of Berk and Jones (1979): For testing a null hypothesis of p-values being the events from a homogeneous Poisson process with constant intensity against an alternative with a single decrease in the intensity, the Berk-Jones generalized likelihood ratio test statistic is

$$\min_{1\le i\le n} i\log(np_{(i)}/i) + (n-i)\log\{n(1-p_{(i)})/(n-i)\}. \tag{12}$$

Starting from (12), Li and Siegmund (2015) proposed the "modified Berk-Jones" (MBJ) statistic

$$\mathrm{BJ}_n^+ = \min_{1\le i\le n/2:\, p_{(i)}<i/n} \{i\log(np_{(i)}/i) + i - np_{(i)}\} \tag{13}$$

to place emphasis on the smaller order statistics. The statistic (13) will also be included in the comparisons.

The first column of Figure 1 shows surface plots for the significance level obtained when combining two p-values using higher criticism, Simes' method, TPM and MBJ. The second column plots the significance level obtained when an additional p-value p is appended to a small list of p-values $(0.05, 0.2, 0.4, 0.8)$ before combining; these plots can be viewed as slices from corresponding five-dimensional surface plots for combining five p-values. For each method, the surfaces are comprised of functions which alternate between constant regions and positive slopes in each p-value axis. Consequently, decreasing any p-value will sometimes not affect the overall significance level, even when this is the smallest p-value. Intuitively, this does not seem satisfactory.

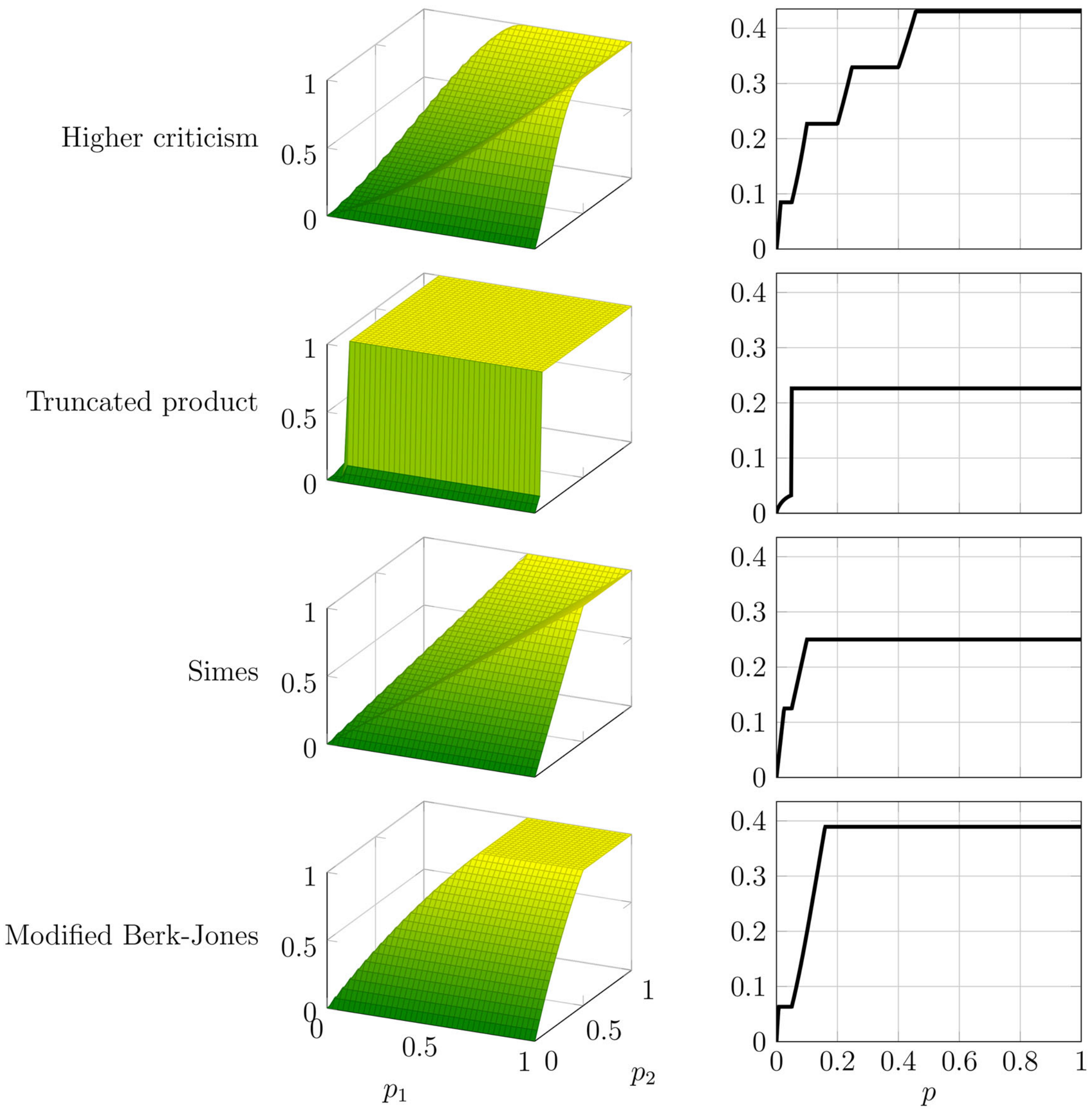

Figure 1. Significance levels from HC, TPM, Simes, MBJ when combining two p-values p_1, p_2 (left), or for five example p-values $(0.05, 0.2, 0.4, 0.8, p)$ as p varies (right).

2. Standardized Partial Sums and Products

This section will give useful theory for understanding the primitive statistics (3) for combining the k smallest of n p-values, and provide motivation for proposing the corresponding statistics (4) for combining p-values in the presence of a sparse signal. Note that the statistics (3) are all nonnegative, but small p-values imply smaller values of $\bar{s}_k^n$ and $\tilde{s}_k^n$ but larger values of s_k^n.

At one extreme $k = n$, s_n^n is Fisher's method for combining p-values, $\tilde{s}_n^n$ is Pearson's method (Pearson 1933) and $\bar{s}_n^n$ is Edgington's method (Edgington 1972). At the other extreme of $k = 1$, s_1^n, $\tilde{s}_1^n$ and $\bar{s}_1^n$ are all equivalent to Tippett's method (Tippett 1931), where the test statistic is the minimum p-value.

These particular cases have simple closed-form null distributions: for $k = n$, $s_n^n, \tilde{s}_n^n \sim \Gamma(n, 1)$ and $\bar{s}_n^n$ has an Irwin–Hall distribution (Hall 1927); for $k = 1, \bar{s}_1^n, \exp(-s_1^n), 1-\exp(-\tilde{s}_1^n) \sim \text{Beta}(1, n)$. In the next section, distributional forms are derived for the general cases of $k < n$.

2.1. Distribution Functions

Proposition 1. Under H_0, for $k < n$, s_k^n has distribution function

$$F_k^n(s) = \sum_{i=1}^{n-k} \left(\frac{k}{i}\right)^{k-1} \frac{(-1)^{k-i}\, n!}{i!k!(n-k-i)!} \left\{ \frac{k(e^{-(k+i)s/k} - 1)}{k+i} + \sum_{j=0}^{k-1} \left(\frac{-i}{k}\right)^j \frac{\gamma(j+1, s)}{j!} \right\}, \quad (14)$$

where γ is the lower incomplete gamma function.

Proof. The quantity s_k^n can be expressed as follows:

$$s_k^n = -k \log p_{(k+1)} - \sum_{i=1}^{k} \log\{p_{(i)}/p_{(k+1)}\}.$$

Under (1), independently $p_{(k+1)} \sim \text{Beta}(k+1, n-k)$ and $p_{(i)}/p_{(k+1)}$, $i = 1, \ldots, k$ are the order statistics of k U[0, 1] random variables. It follows that $s_k^n \overset{d}{\equiv} X + X'$ where $e^{-X/k} \sim$

$\text{Beta}(k+1, n-k)$ and $X' \sim \Gamma(k, 1)$. Noting that X would have density

$$f_X(x) = \frac{n!}{k!} \sum_{i=1}^{n-k} \frac{i}{k} \frac{(-1)^{i+1}}{i!(n-k-i)!} e^{-(k+i)x/k},$$

the result follows from calculating the convolution with the gamma distribution function. □

Proposition 2. Under H_0, for $k < n$, $\tilde{s}_k^n$ has distribution function

$$\tilde{F}_k^n(s) = \sum_{i=1}^{k} \frac{(-1)^{k-i}\, n!}{(i-1)!(n-k)!(k-i)!(n-k+i)} \left(\frac{i}{n-k}\right)^{k-1} \left(1 - e^{-(n-k+i)s/i}\right). \tag{15}$$

Proof. Note that $\tilde{s}_k^n = -\sum_{i=1}^{k} \log(1-p_{(i)}) \overset{d}{\equiv} -\sum_{i=n-k+1}^{n} \log p_{(i)}$, and thus equal in distribution to the sum of the first k order statistics of n independent standard exponential random variables. Hence, $\tilde{s}_k^n \overset{d}{\equiv} \sum_{i=1}^{k} X_i$, where $X_i \sim \exp(1 + (n-k)/i)$ and $\tilde{s}_k^n$ has a hypoexponential distribution. □

Proposition 3. Under H_0, for $k < n$, $\bar{s}_k^n$ has distribution function

$$\bar{F}_k^n(s) = \sum_{i=0}^{k} \frac{(-1)^i\, n!}{i!(k-i)!(n-k-1)!} \sum_{j=0}^{k} \frac{(s-i)^{k-j} i^j \left(1 - \{1 - \min(s/i, 1)\}^{j+n-k}\right)}{j!(k-j)!(j+n-k)}, \tag{16}$$

Proof. The result follows from $\bar{s}_k^n$ being the product of a $\text{Beta}(k+1, n-k)$ variable and an Irwin–Hall distribution from summing k independent standard uniform random variables. □

The corresponding density functions derived from these three distribution functions are illustrated in Figure 2, for the case $n = 10$ and different values of k. There is clear right-skew for $k = 1$, but for increasing k the densities quickly become more symmetric.

Unfortunately, for $n > 13$, Equations (14)–(16) are all numerically unstable, with the alternating signed terms in each equation leading to *catastrophic cancellation*. This problem is illustrated in Figure 3, which plots on the log-scale, for two example values of k, the rapid divergence of the smallest and largest absolute values of the terms in each of the distribution functions (14)–(16) as n increases; for each (k, n) pair, the summands of each function are evaluated at the corresponding distribution expected values, which will be derived in Section 2.2. Note the scale of magnitude of the extreme values in Figure 3, with the resulting sum still lying in $[0, 1]$.

2.2. Central Moments

In contrast to the numerical instability of the distributions for each primitive statistic s_k^n, $\tilde{s}_k^n$, and $\bar{s}_k^n$, analytic expressions are available for the means and variances which can be reliably evaluated for any values of n and k. The moment equations for s_k^n and $\tilde{s}_k^n$ follow directly from the distributions and decompositions used in Propositions 1 and 2.

Proposition 4. Suppose H_0. For $1 \le k \le n$, the sum of p-values $\bar{s}_k^n$ has mean and variance

$$\begin{aligned} \mathbb{E}(\bar{s}_k^n) &= k(k+1)/(2(n+1)), \\ \mathbb{V}(\bar{s}_k^n) &= k(k+1)\{2n(1+2k) - (k-1)(3k+2)\}/ \\ &\qquad \{12(n+1)^2(n+2)\}. \end{aligned}$$

For the sum of log p-values, s_k^n,

$$\begin{aligned} \mathbb{E}(s_k^n) &= k\{1 - \psi(k+1) + \psi(n+1)\}, \\ \mathbb{V}(s_k^n) &= k + k^2\{\psi_1(k+1) - \psi_1(n+1)\}, \end{aligned} \tag{17}$$

where ψ and ψ_1, respectively, denote the digamma and trigamma functions. For the complementary sum, $\tilde{s}_k^n$,

$$\begin{aligned} \mathbb{E}(\tilde{s}_k^n) &= k\{\psi(k+1) - \psi(n+1)\}, \\ \mathbb{V}(\tilde{s}_k^n) &= k + (n-k)^2\{\psi_1(n-k+1) - \psi_1(n+1)\} \\ &\qquad -2(n-k)\{\psi(n+1) - \psi(n-k+1)\}. \end{aligned}$$

2.3. Gaussian Approximation

The numerical instability of the distributional formulae for $s_k^n, \tilde{s}_k^n, \bar{s}_k^n$ as n and k increase suggests numerically stable approximations would be valuable for these cases. The following results provide central limit theorem (CLT) approximations for all three statistics, using the central moments from Proposition 4.

Theorem 1. As $k, n \to \infty$, with k growing sufficiently fast with n, for $s \in \{s_k^n, \tilde{s}_k^n, \bar{s}_k^n\}$,

$$\{s - \mathbb{E}(s)\}/\sqrt{\mathbb{V}(s)} \xrightarrow{d} \mathrm{N}(0, 1).$$

Proof. An alternative (hypoexponential) representation for the null probability distribution of $s_k^n = -\sum_{i=1}^{k} \log p_{(i)}$ is $s_k^n \overset{d}{\equiv} \sum_{i=1}^{n} X_i$ where the variables $X_, \ldots, X_n$ are independent, exponentially distributed variables with corresponding rate parameters $\lambda_1, \ldots, \lambda_n$ satisfying $\lambda_i = \max\{1, \frac{i}{k}\}$. The means and variances of these random variables are, respectively, $\mu_i = \lambda_i^{-1}$ and $\sigma_i^2 = \lambda_i^{-2}$. Defining $\xi_{n,k} = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$, then $\max_{1 \le i \le n} \sigma_i^2 = 1$ and $\xi_{n,k}^2 = k + k^2\{\psi_1(k+1) - \psi_1(n+1)\}$ from (17). Since $k < \xi_{n,k}^2 \le n$,

$$\frac{\max_{1 \le i \le n} \sigma_i^2}{\xi_{n,k}^2} \to 0,$$

as $k, n \to \infty$. Appealing to the Lindeberg central limit theorem, for $\epsilon > 0$,

$$\begin{aligned} &\textstyle\sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mu_i)^2 \mathbb{1}_{\{x:|x-\mu_i|>\epsilon\,\xi_{n,k}\}}(X_i)\right] = \\ &\textstyle\sum_{i=1}^{n} \left\{ I(\lambda_i, x)\Big|_{x=0}^{x=\max\{0, \lambda_i^{-1} - \epsilon\,\xi_{n,k}\}} - I\left(\lambda_i, \lambda_i^{-1} + \epsilon\,\xi_{n,k}\right) \right\} \end{aligned}$$

where $I(\lambda, x) = -e^{-\lambda x}(x^2 + \lambda^{-2})$ and for all i, $\lambda_i \ge 1$. As $k, n \to \infty$, the widths of intervals in the left-hand terms all shrink to exactly zero, and for the right-hand terms,

$$\begin{aligned} &\textstyle\sum_{i=1}^{n} -I\left(\lambda_i, \lambda_i^{-1} + \epsilon \lambda_i \xi_{n,k}\right) \\ &\quad = \textstyle\sum_{i=1}^{n} e^{-(1+\epsilon\,\xi_{n,k})} \left\{ \left(\lambda_i^{-1} + \epsilon\,\xi_{n,k}\right)^2 + \lambda_i^{-2} \right\} \\ &\quad \le n\, e^{-(1+\epsilon\,\xi_{n,k})}\{(1 + \epsilon\,\xi_{n,k})^2 + 1) \\ &\quad < n\, e^{-(1+\epsilon\sqrt{k})}\{(1 + \epsilon n)^2 + 1\} \\ &\quad \to 0 \end{aligned}$$

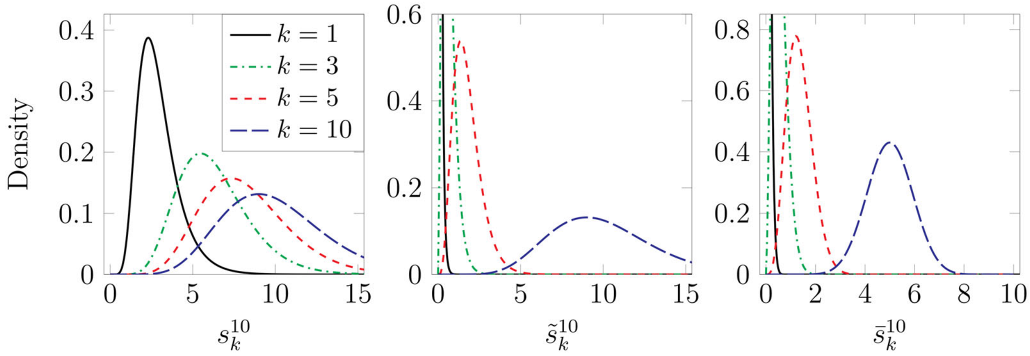

Figure 2. Null probability density functions of s_k^n, $\tilde{s}_k^n$ and $\bar{s}_k^n$ for $n = 10$ and $k = 1, 3, 5, 10$.

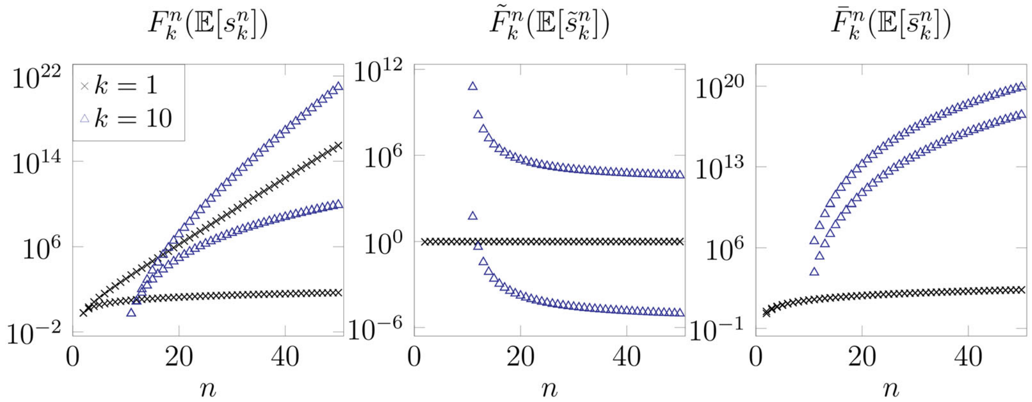

Figure 3. The smallest and largest absolute value terms in the outer summations of (14), (15), and (16) when evaluated at the corresponding distribution mean values.

as $n \to \infty$ if $\sqrt{k} \in \omega(\log n)$. Then

$$\frac{1}{\sum_{i=1}^n \sigma_i^2} \sum_{i=1}^n \mathbb{E}\left[(X_i - \mu_i)^2 \mathbb{1}_{\{x:|x-\mu_i|>\epsilon\ \xi_{n,k}\}}(X_i)\right] \to 0.$$

As noted in the proof to Proposition 2, $\tilde{s}_k^n = -\sum_{i=1}^k \log(1 - p_{(i)})$ follows a hypoexponential distribution as the sum of k independent exponential random variables with the ith rate parameter $\lambda_i = \frac{n-k+i}{i} \geq 1$, and the result follows analogously to the proof for s_k^n.

Finally, $\bar{s}_k^n = \sum_{i=1}^k p_{(i)}$ is equivalent to a sum of k exchangeable random variables, $\bar{s}_k^n = \sum_{i=1}^k X_i$ where independently $X_i \sim \mathrm{U}(0, X')$, $i = 1, \ldots, k$ and $X' \sim \mathrm{Beta}(k+1, n-k)$. The CLT of Blum et al. (1958) for sequences of exchangeable processes provides the following, sufficient conditions, the covariances:

$$\mathrm{cov}(X_1, X_2) = \frac{(k+1)(n-k)}{4(n+1)^2(n+2)},$$
$$\mathrm{cov}(X_1^2, X_2^2) = \frac{2(k+1)(k+2)(n-k)(2kn + 5(n+k) + 11)}{9(n+1)^2(n+2)^2(n+3)(n+4)}$$

must converge to zero, the former at rate $o(1/n)$, and

$$\mathbb{E}(X_1^3) = \frac{(k+1)(k+2)(k+3)}{4(n+1)(n+2)(n+3)}$$

should be $o(\sqrt{n})$. These rates are obtained when $(n-k) \in o(n)$. □

2.4. Standardized Statistics

To construct test statistics based on the partial sums and products $\tilde{s}_k^n, s_k^n, \bar{s}_k^n$ which optimize over k, a natural approach would be to consider, for example,

$$\max_{k=1,\ldots,n} F_k^n(s_k^n) \tag{18}$$

for finding the most significant partial product. However, the numerical instability of the analytic expressions (1)–(3), for even moderate n, invalidates this approach. A straightforward solution would construct Monte Carlo estimates of the distribution functions for each $k < n$, but this becomes too cumbersome for large n.

Instead, following Donoho and Jin (2004), a pragmatic solution is to exploit the CLT results from Theorem 1 and approximate the true distributions (14)–(16) with Gaussians, yielding the simple standardized test statistics PP, PCP, and PS (4) introduced in Section 1; note that if the true distributions were Gaussian, the two definitions (18) and (4) coincide

$$\max_{k=1,\ldots,n} F_k^n(s_k^n) \approx \max_{k=1,\ldots,n} \Phi\left(\{s_k^n - \mathbb{E}(s_k^n)\}/\sqrt{\mathbb{V}(s_k^n)}\right)$$
$$= \Phi\left(\max_{k=1,\ldots,n} \{s_k^n - \mathbb{E}(s_k^n)\}/\sqrt{\mathbb{V}(s_k^n)}\right).$$

From Figure 2, similarly to higher criticism (5), the Gaussian approximation of normality is seen to be poor for small k due to the positive skew, but quickly improves as k increases. Furthermore, the standardized statistics will be seen to perform well in practice.

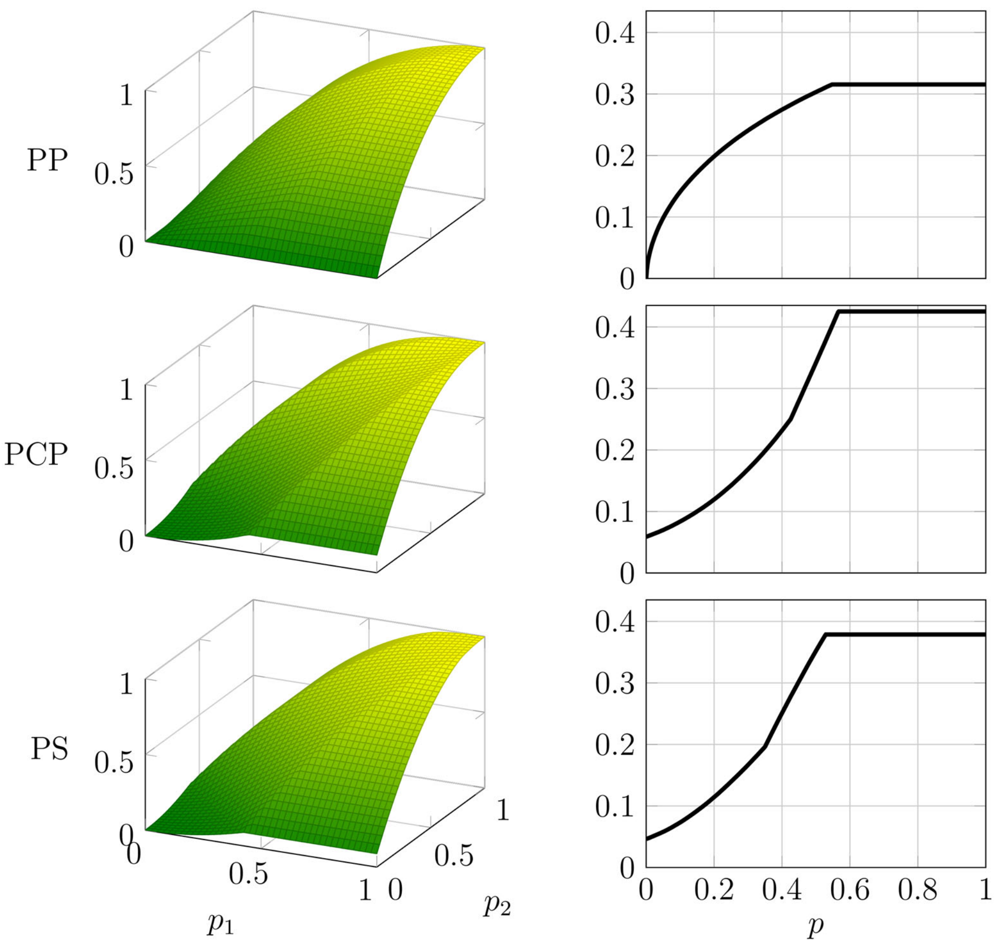

Figure 4. Significance levels from PP, PCP, PS (4) when combining two p-values p_1, p_2 (left), or for five example p-values $(0.05, 0.2, 0.4, 0.8, p)$ as p varies (right).

The distributions of the statistics (4) are comparatively simple to obtain by Monte Carlo simulation, providing the necessary guarantee of a well calibrated test under H_0. Figure 4 provides analogous significance level plots to those from Figure 1, obtained for the test statistics PP, PCP, and PS. In contrast to Figure 1, these statistics (in particular PP) display smooth changes in the resulting significance level as individual p-values are varied. Like the original statistics of Fisher (1929), Pearson (1933), and Edgington (1972) without partial sums, only PP gives a combined significance level of zero when just one p-value is zero.

2.5. Related Literature

Despite the partial products and sums (3) being very natural statistics for combining the k smallest p-values, the closed-form expressions for their distribution functions (14)–(16) have not been found in searches of the existing literature on combining p-values, and neither have the central limit theorem Gaussian approximations. Interestingly, one of the three linear standardization techniques (4), the partial product, has appeared in the context of genomic analyses, with PP_n used to determine copy number variations in DNA sequences (Song, Min, and Zhang 2016); however, the theoretical justifications for using this statistic, as well as the extensions to partial complementary products and partial sums, are novel.

In the mainstream p-value combination literature, two main alternatives have been considered: The first simply fixes the number of p-value order statistics to be combined, providing the simple "rank truncated product" statistic of Dudbridge and Koeleman (2003), $W_R = \prod_{i=1}^{k} p_{(i)}$, for a fixed integer $k < n$. Second, more satisfactory approaches have tried to optimize over k with respect to the corresponding null distribution functions (18) using, for each nontrivial choice of $1 < k < n$, a permutation testing procedure (Yu et al. 2009; Li and Tseng 2011) or Monte Carlo simulation or other numerical integration techniques (Zhang, Chen, and Pfeiffer 2013). However, while these techniques promise an approximation to the desired optimization of Equation (3) over k, avoiding the numerical instability of the closed-form analytic expressions for these distributions, their Monte Carlo storage requirement scales linearly with n; for M Monte Carlo samples of n uniform p-values, an $M \times (n-2)$ matrix of the primitive statistic (3) for each sample, for each $1 < k < n$ is required; the columns provide empirical distribution estimates for each k, which are then applied the those columns (essentially ranking the entries), before being minimized for each row and to provide a doubly Monte Carlo estimate. This makes these latter techniques unsuitable for the large-scale testing problems considered here, and also unsuited to widespread usage. Huo et al. (2020) recently released an R package, *AWFisher*, for combined p-values from (18) by interpolating from stored look-up tables obtained via importance sampling, but again only allowing samples sizes up to a maximum of $n = 100$.

2.6. Computation

In the simulations in the next section, the Monte Carlo scheme described above is deployed with 100,000 samples for values of n up to 10,000 and included in those comparisons; for larger n, storage was not feasible for a computer with 16GB memory.

To calculate the null distributions of the test statistics (4), for each value of n considered, PP, PCP, and PS, were calculated for each of 1 million samples of n independent uniform variables, providing Monte Carlo estimates of the distributions which in turn provide corresponding p-values for an observed sample of n p-values. It should be noted that although the standardized statistics were motivated by asymptotic normality results, the Monte Carlo estimated distributions do not rely upon these properties and are unbiased. Consequently, as with all of the methods compared in this article, all calculated significance levels are theoretically exact (using either analytic distribution functions or high-precision Monte Carlo estimates) and therefore control of Type I error rates is guaranteed.

3. Power Comparisons

Standardized partial sums and products (4) are now empirically compared with higher criticism (5) and (9), Simes' method (10), the truncated product method (11), the modified Berk-Jones method (13) and, where numerically feasible, Monte Carlo simulation for partial products (18). The methods are first compared on synthetic p-values from tests of mixtures of Gaussians (6), and second on p-values generated from a real data analysis in model-testing for computer network cyber-security.

3.1. Mixture of Gaussians

Recall the Asymptotic Rare/Weak model (6) of n unit variance Gaussian variables, where a diminishing in n proportion $\epsilon_n = n^{-\beta}$ have nonzero mean $\mu_n = \sqrt{2r \log n}$; Donoho and Jin (2004) showed higher criticism asymptotically dominated Fisher's method for combining p-values under this model for a range of (β, r) values $\beta \in (0.5, 1)$ and $r > \rho^*(\beta)$. Picking values $\beta = 2/3$ and $r = \rho^*(2/3) + 0.1 = 4/15$ well within this region, Figure 5 shows the distribution of combined p-values obtained for each method as n is increased by factors of 100. In each plot, the curve for a particular method shows the probability under the alternative hypothesis of obtaining a combined p-value not exceeding p.

Initially, there is little to distinguish the different methods, but by $n = 1{,}000{,}000$ the asymptotic optimality of HC over Fisher's method is apparent. However, the strongest performance is achieved by the standardized statistics (4) and in particular PP, which is most striking when the plot is zoomed in to the more interesting range of low combined p-values in the bottom-right quadrant of Figure 5 through use of a log-scale.

The dotted line referred to as "MC partial product" corresponds to Monte Carlo estimation of the distribution function (14) for each $k < n$; pleasingly, this gives visually indistinguishable performance to the corresponding standardization technique PP. However, for n larger than 10,000 the MC method could no longer be feasibly deployed.

The lower dashed line in the bottom right panel of Figure 5 also shows the power curve for the modified higher criticism statistic HC^+ (9) for $n = 1{,}000{,}000$. This outperforms HC toward the interesting end of producing very low combined p-values. Similarly, a small uplift can be obtained under this model by modifying the best standardized statistic PP,

$$\text{PP}_n^+ = \min_{1 \le k \le n : p_{(k)} \ge 1/n} \{\mathbb{E}(s_k^n) - s_k^n\} / \sqrt{\mathbb{V}(s_k^n)}, \tag{19}$$

which is represented in Figure 5 by the longer-dashed line. Note that Simes' method is not competitive, and the truncated product method is close in performance to Fisher's method. The modified Berk-Jones statistic proposed by Li and Siegmund (2015) outperforms both versions of higher criticism, but is below the standardized partial product.

Figure 6 shows the distribution of p-values under H_1 for different values of the parameter β from (8) which controls the proportion of alternative hypothesis p-values in the sparse H_1 signal; for each value of β, the effect size parameter r from (7) was kept at $\rho^*(\beta) + 0.1$ to provide adequate separation of the rival methods. The sample size n was kept to $10,000$ to accommodate the high storage requirement of Monte Carlo estimation procedure referred to as MC partial product.

In all cases, the MC partial product tracks the corresponding standardization technique PP fairly closely, and these two partial product techniques are clearly the best performing across the range of β values. The modified Berk–Jones statistic outperforms higher criticism for smaller values of β, but under performs when β is increased, corresponding to a sparser signal. For $\beta = 0.9$, it is expected that only two or three of the 10,000 p-values will be from the alternative density, and the probability of getting small combined p-values is seen to be very similar for Simes' method, HC and PP.

3.2. A Closer Examination of Higher Criticism

To understand why the standardized partial product statistics are outperforming higher criticism under the Asymptotic Rare/Weak model, it is useful to find other examples where higher criticism is more powerful. To see how to construct such an example, clues can be obtained from the theoretical significance curves from Figure 1. It was remarked in Section 1 that small changes to even the smallest p-values do not effect higher criticism or similar methods; for higher criticism to be powerful, this weakness needs to be of limited importance. To this end, consider a modified alternative hypothesis for mixtures of Gaussians:

$$H_{1,i}: \quad t_i \sim \text{N}(\mu_n, \sigma_n^2), \tag{20}$$

where $\sigma_n^2 \ll 1$, but still using p-values $p_i = 1 - \Phi(t_i)$: Crucially under this formulation, since the rare draws from the alternative Gaussian have very low variance, then all of the significant p-values will be similar to one another.

It should be noted this example is presented with some discomfort. Usually in studies of combining p-values, the alternative density of p-values is assumed non-increasing on $[0, 1]$,

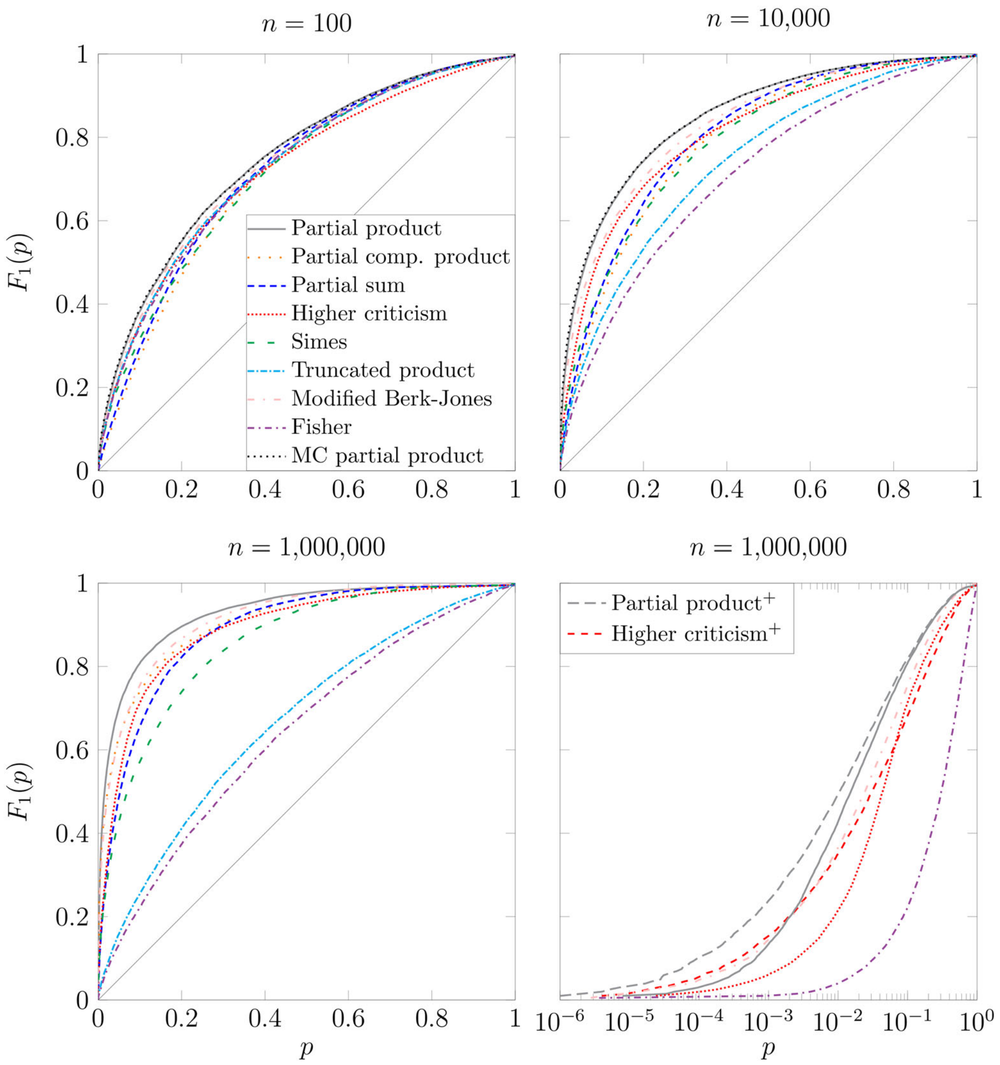

Figure 5. Distribution of significance levels from combining n p-values under the mixture of Gaussians (6). The thin identity line corresponds to H_0 and H_1 being indistinguishable.

so that only combination methods which are monotonic in the p-values are admissible (Birnbaum 1954). However, if $t \sim \mathrm{N}(\mu, \sigma^2)$, then the p-values $p = 1 - \Phi(t)$ have density

$$f_1(p) = \exp[-\{\Phi^{-1}(1-p) - \mu\}^2/(2\sigma^2) + \Phi^{-1}(1-p)^2/2]/\sigma, \quad (21)$$

which is not monotonic for $\sigma \neq 1$. As a consequence, there is a critical value

$$p^*(\mu,\sigma) = \Phi[\{\mu + \sigma\sqrt{\mu^2 + 2(\sigma^2-1)\log\sigma}\}/(\sigma^2-1)],$$

such that for $p < p^*(\mu,\sigma) \implies f_1(p) < 1 = f_0(p)$. For $\mu = 2, \sigma = 0.05$, this translates to $p^*(2, 0.05) \approx 0.0153$, and so p-values lower than 1%, for example, are more probable under H_0 than H_1. The mixture densities for the test statistics and their p-values (21) under (20) are illustrated in Figure 7 with $\mu = 2, \sigma = 0.05, \epsilon = 1/25$.

For comparison with Figure 5, Figure 8 shows the distribution of significance levels when combining p-values from this mixture of Gaussian distributions with $\sigma_n = 0.05$, $\beta = 2/3$ and $r = \rho^*(2/3) + 0.25 = 5/12$. The truncated product method performs best for small n, but as n increases, higher criticism and in particular the modified version HC$^+$ is superior. Asymptotically, it seems Simes' method cannot distinguish between H_0 and H_1 here.

3.3. Computer Network Modeling

The methods are now compared using p-values obtained from real data on the waiting times between different authentication event types in a computer network, previously analyzed in Price-Williams, Heard, and Rubin-Delanchy (2019). The aim of that study was to assess temporal causality between different types of computer network authentication. Learning dependencies in network events is an important step in building realistic probability models of computer networks for cyber-security statistical anomaly detection. In particular, focus was on detecting *weak dependencies* where only a small subset of events were causal; the illustrative example examined screensaver dismissals leading to a wrong password failed authentication. For each user, Price-Williams, Heard, and Rubin-Delanchy (2019) obtained the lower tail p-value from a fitted Hawkes process model for each waiting time (between dismissal and failed password), and combined these p-values to give an overall significance level

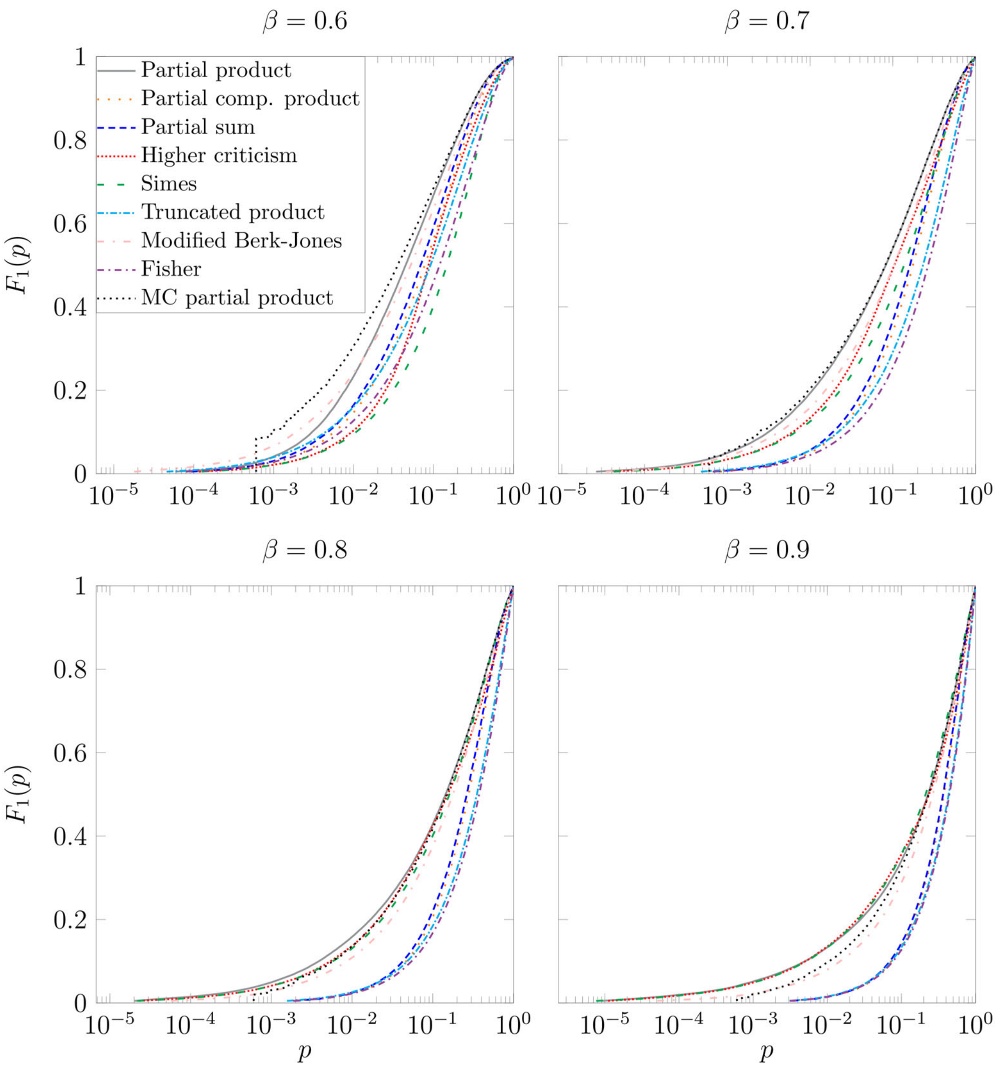

Figure 6. Distribution of significance levels from combining $n = 10{,}000$ p-values under the mixture of Gaussians (6) with different values of β and $r = \rho^*(\beta) + 0.1$.

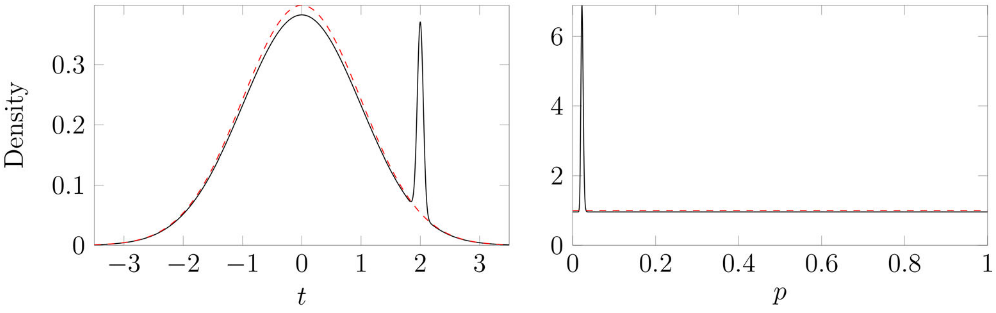

Figure 7. Probability densities of Gaussian mixture test statistics (left) and the corresponding p-values (right) under the null (red, dashed line) and alternative (black, solid line) hypotheses.

for that user. A comparison of methods there showed modified higher criticism HC^+ strongly outperformed a similarly modified Fisher's method across users in detecting dependence due to the sparsity in the signal, since most screensaver dismissals will be followed by a successful password entry.

Starting with the same sequences of p-values for each user, the left panel of Figure 9 shows the distribution of significance levels obtained from each method. The figure only shows combined p-values up to a threshold of 0.01, since most methods eventually detect significance at higher false-positive rates. The right panel shows how frequently over 1051 users each method yielded the smallest combined p-value. Note that due to finite Monte Carlo sampling (10^8 null samples), some of the methods were tied in their performance for some users (particularly with estimated p-values of 0); for a fair comparison, p-values from methods which have an analytic distribution were rounded to the same level of precision.

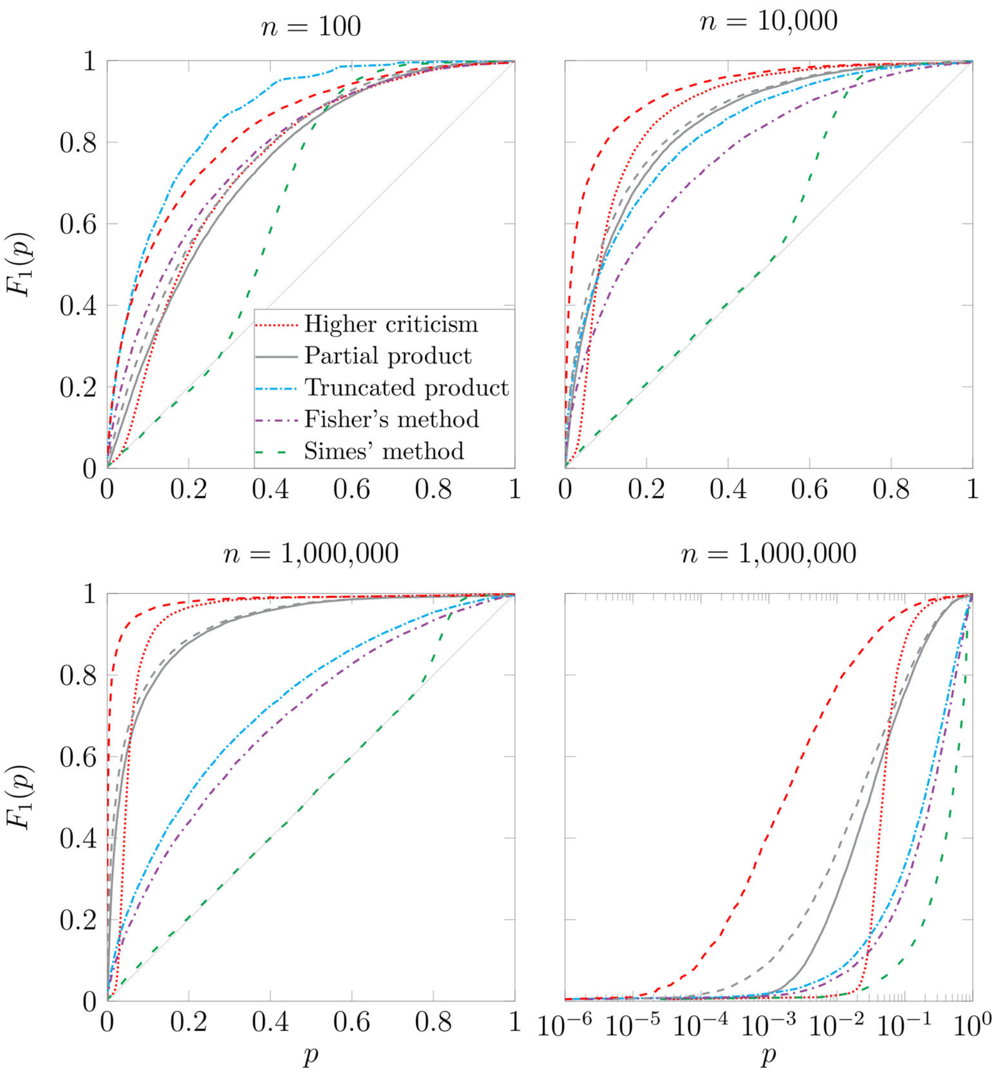

Figure 8. Distribution of significance levels from combining n p-values under the revised mixture of Gaussians from H_1 (20) and Figure 7, well-suited to the higher criticism statistic.

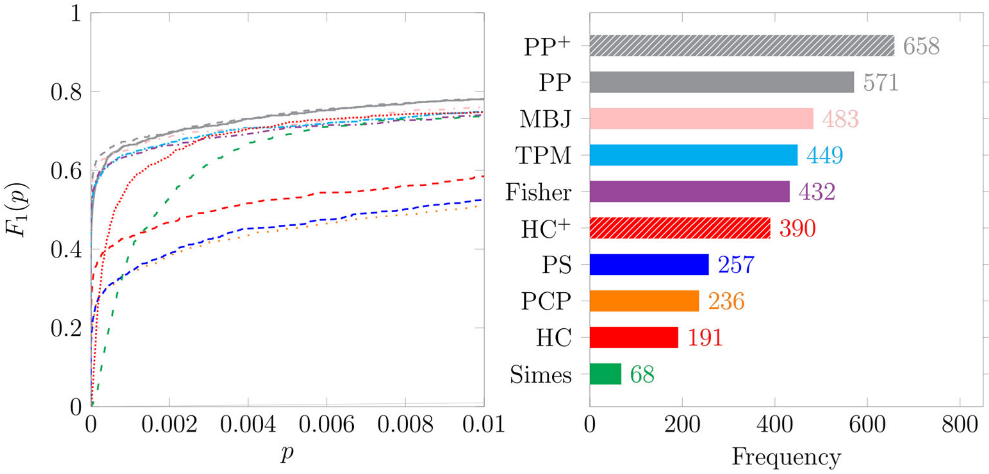

Figure 9. Meta-analysis of dependency in computer network traffic. Left: The distribution of significance levels of different p-values combining methods. Right: The number of cases for which each method yielded the lowest combined p-value.

The partial product, and in particular the modified version (19), are the best performing methods; higher criticism is outperformed by the modified Berk-Jones and truncated product methods, and Simes' method is disappointingly weak. To quantify performance, Table 1 shows partial areas under curves, restricted to small p-values up to 0.01 (McClish 1989).

Table 1. Partial AUCs for *p*-value combination methods on computer network data

Method	PP$^+$	PP	MBJ	TPM	Fisher	HC	Simes	HC$^+$	PS	PCP
AUC	0:733	0:729	0:712	0:701	0:694	0:679	0:618	0:508	0:447	0:437

The partial product provides a robust, powerful combination method.

4. Discussion

A closed-form expression for the null distribution, F_k^n of the sum of logs, s_k^n, of the k smallest of n standard uniform p-values has been derived, but shown to be numerically unstable in practice. However, motivated by a central limit theorem result, standardizations of s_k^n have been empirically shown in practice to behave almost indistinguishably from computationally expensive Monte Carlo estimates of F_k^n. Analogous results have been presented for the sums of complementary logs and untransformed p-values, $\tilde{s}_k^n$ and $\bar{s}_k^n$, presenting a useful triple of related methods.

The standardized partial product, which extends the method of Fisher (1929), has been shown to provide higher power for testing p-values arising from the canonical mixture of Gaussians used for illustrating the higher criticism (Donoho and Jin 2004) and modified Berk–Jones (Li and Siegmund 2015) statistics, which each offer asymptotic dominance over Fisher's method. Although obtaining formal proof that the proposed methods reach the same asymptotic detection boundary is an open problem, the empirical comparisons, where the standardized partial product is shown to dominate higher criticism as n reaches ten thousand and then one million, strongly suggest that the same asymptotic property should hold for this and many of the other methods, with the possible exception of the truncated product. The standardized partial product was also demonstrated to be the most powerful combiner among those compared in a practical computer network modeling example.

Supplementary Material

Python code: Code implementing all of the *p*-value combination methods can be obtained from *https://github.com/naheard/standardised_partial_product.git*.

References

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [2]

Berk, R. H., and Jones, D. H. (1979), "Goodness-of-Fit Test Statistics That Dominate the Kolmogorov Statistics," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 47, 47–59. [2]

Birnbaum, A. (1954), "Combining Independent Tests of Significance," *Journal of the American Statistical Association*, 49, 559–574. [1,8]

Blum, J. R., Chernoff, H., Rosenblatt, M., and Teicher, H. (1958), "Central Limit Theorems for Interchangeable Processes," *Canadian Journal of Mathematics*, 10, 222–229. [5]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *Annals of Statistics*, 32, 962–994. [2,5,7,11]

——— (2015), "Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects," *Statistical Science*, 30, 1–25. [2]

Dudbridge, F., and Koeleman, B. P. (2003), "Rank Truncated Product of p-Values, With Application to Genomewide Association Scans," *Genetic Epidemiology*, 25, 360–366. [6]

Edgington, E. S. (1972), "An Additive Method for Combining Probability Values From Independent Experiments," *The Journal of Psychology*, 80, 351–363. [3,6]

Fisher, R. A. (1929), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd. [1,6,11]

Hall, P. (1927), "The Distribution of Means for Samples of Size n Drawn From a Population in Which the Variate Takes Values Between 0 and 1, All Such Values Being Equally Probable," *Biometrika*, 19, 240–245. [3]

Heard, N. A., and Rubin-Delanchy, P. (2018), "Choosing Between Methods of Combining *p*-Values," *Biometrika*, 105, 239–246. [1]

Huo, Z., Tang, S., Park, Y., and Tseng, G. (2020), "P-Value Evaluation, Variability Index and Biomarker Categorization for Adaptively Weighted Fisher's Meta-Analysis Method in Omics Applications," *Bioinformatics*, 36, 524–532. [6]

Li, J., and Siegmund, D. (2015), "Higher Criticism: *p*-Values and Criticism," *Annals of Statistics*, 43, 1323–1350. [2,7,11]

Li, J., and Tseng, G. C. (2011), "An Adaptively Weighted Statistic for Detecting Differential Gene Expression When Combining Multiple Transcriptomic Studies," *Annals of Applied Statistics*, 5, 994–1019. [6]

McClish, D. K. (1989), "Analyzing a Portion of the ROC Curve," *Medical Decision Making*, 9, 190–195. [10]

Pearson, K. (1933), "On a Method of Determining Whether a Sample of Size *n* Supposed to Have Been Drawn From a Parent Population Having a Known Probability Integral Has Probably Been Drawn at Random," *Biometrika*, 25, 379–410. [3,6]

Price-Williams, M., Heard, N., and Rubin-Delanchy, P. (2019), "Detecting Weak Dependence in Computer Network Traffic Patterns by Using Higher Criticism," *Journal of the Royal Statistical Society*, Series C, 68, 641–655. [8]

Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754. [2]

Song, C., Min, X., and Zhang, H. (2016), "The Screening and Ranking Algorithm for Change-Points Detection in Multiple Samples," *Annals of Applied Statistics*, 10, 2102–2129. [6]

Tippett, L. H. C. (1931), *The Methods of Statistics*, London: Williams and Norgate, Ltd. [3]

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009), "Pathway Analysis by Adaptive Combination of p-Values," *Genetic Epidemiology*, 33, 700–709. [6]

Zaykin, D., Zhivotovsky, L. A., Westfall, P., and Weir, B. (2002), "Truncated Product Method for Combining p-Values," *Genetic Epidemiology*, 22, 170–185. [2]

Zhang, S., Chen, H., and Pfeiffer, R. (2013), "A Combined p-Value Test for Multiple Hypothesis Testing," *Journal of Statistical Planning and Inference*, 143, 764–770. [6]