A powerful goodness-of-fit test using the probability integral transform of order statistics

By Christian T. Covington

Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

ccovington@g.harvard.edu

Jeffrey W. Miller

Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

jwmiller@hsph.harvard.edu

Summary

Goodness-of-fit (GoF) tests are a fundamental component of statistical practice, essential for checking model assumptions and testing scientific hypotheses. Despite their widespread use, popular GoF tests exhibit surprisingly low statistical power against substantial departures from the null hypothesis. To address this, we introduce PITOS, a novel GoF test based on applying the probability integral transform (PIT) to the jth order statistic (OS) given the ith order statistic for selected pairs i, j. Under the null, for any pair i, j, this yields a Uniform$(0, 1)$ random variable, which we map to a p-value via $u \mapsto 2\min(u, 1-u)$. We compute these p-values for a structured collection of pairs i, j generated via a discretized transformed Halton sequence, and aggregate them using the Cauchy combination technique to obtain the PITOS p-value. Our method maintains approximately valid Type I error control, has an efficient $O(n \log n)$ runtime, and can be used with any null distribution via the Rosenblatt transform. In empirical demonstrations, we find that PITOS has much higher power than popular GoF tests on distributions characterized by local departures from the null, while maintaining competitive power across all distributions tested.

Some key words: goodness-of-fit; hypothesis test; nonparametric test; test of normality; test of uniformity

1. Introduction

Goodness-of-fit (GoF) testing is a key part of the statistical toolkit, widely used for validating modeling assumptions and testing scientific hypotheses. By design, GoF tests typically assess the fit of observed data to a null distribution without requiring a pre-specified alternative. This generality makes GoF tests very flexible, but also means that they are often applied with minimal regard to the types of deviations from the null that might occur. As such, ideally a GoF test would have high power to detect a wide range of deviations from the null (D'Agostino & Stephens, 1986; Lehmann & Romano, 2005).

However, the most popular GoF tests often have surprisingly low statistical power, even on distributions that are conspicuously different from the null. To illustrate, we compute the power of four popular GoF tests on data from the six distributions shown in Figure 1, taking Uniform$(0, 1)$

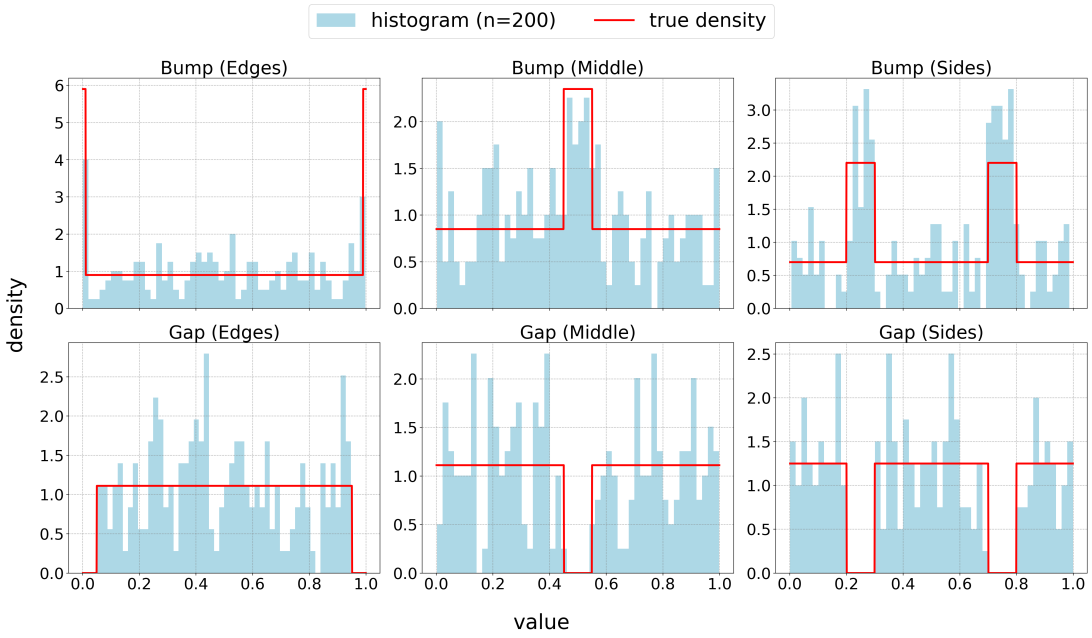

Fig. 1: Examples of data distributions on which popular GoF tests have low power. For each, the true density is shown in red, and a histogram of one simulated data set is shown in blue.

Test	Bump (Edges)	Gap (Edges)	Bump (Middle)	Gap (Middle)	Bump (Sides)	Gap (Sides)
Anderson–Darling	0.843	0.529	0.387	0.205	0.392	0.256
Neyman–Barton	0.765	0.810	0.540	0.327	0.114	0.152
Kolmogorov–Smirnov	0.189	0.114	0.593	0.376	0.516	0.387
Cramér–von Mises	0.216	0.152	0.428	0.197	0.257	0.154
LRT (Oracle)	1.0	1.0	1.0	1.0	1.0	1.0

Table 1: Power of commonly used GoF tests on data from the distributions in Figure 1. We approximate the power at the $\alpha = 0.05$ level by averaging over 100,000 simulated data sets with a sample size of $n = 200$. Red text indicates power ≤ 0.25.

to be the null hypothesis. Specifically, we consider the Anderson–Darling (AD), Neyman–Barton (NB), Kolmogorov–Smirnov (KS), and Cramér–von Mises (CvM) tests; see Section S1.

In Table 1, we see that each of these tests exhibits low power on most of these distributions. In particular, KS has very low power in the tails, AD has very low power around the median, NB has very low power around the first and third quartiles, and CvM has very low power across the board. This poor performance is not because these departures from the null are intrinsically hard to detect. Indeed, power very close to 1 is obtained by an oracle benchmark based on the likelihood ratio test (LRT) using the true distribution as the alternative hypothesis (Table 1); see Section S1 for details.

We introduce a novel GoF test, referred to as the *Probability Integral Transform of Order Statistics* (PITOS), that has high power in many cases where these popular tests fail and has competitive power across all cases that we have explored. PITOS takes a sequence of pairs of indices $i, j \in \{1, \ldots, n\}$, uses the conditional distribution of the order statistics $X_{(j)} \mid X_{(i)}$ under the null to calculate a p-value p_{ij} for each pair (i, j), and combines these p-values via Cauchy

combination to get a single combined p-value. This approach admits a large class of tests via the choice of pairs (i, j), so we propose a default which generates pairs of indices via a discretized transformed Halton sequence. We show that PITOS maintains approximately valid Type I error control and has an $O(n \log n)$ runtime. We perform simulation studies comparing the power of PITOS and other GoF tests on a range of distributions.

2. Methodology

In this section, we introduce our proposed goodness-of-fit test, the *Probability Integral Transform of Order Statistics* (PITOS). Given independent and identically distributed (i.i.d.) data $X_1, \ldots, X_n$, we are interested in testing whether we can reject a point null hypothesis, which we take to be that $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. The choice of $\text{Uniform}(0, 1)$ is made without loss of generality since for any null with a continuous cumulative distribution function (CDF) F, if $Y_i \sim F$ then $F(Y_i) \sim \text{Uniform}(0, 1)$; thus, we can first map the data through F and then apply a test based on a uniform null. More generally, any random vector $(Y_1, \ldots, Y_n)$ with a known joint distribution can be mapped to a vector of i.i.d. $\text{Uniform}(0, 1)$ random variables via the generalized Rosenblatt transform (Brockwell, 2007); see Section S2 for details.

2.1. *PITOS goodness-of-fit test*

Let $X_1, \ldots, X_n$ i.i.d. $\sim \text{Uniform}(0, 1)$. The order statistics $X_{(1)}, \ldots, X_{(n)}$ are defined by arranging $X_1, \ldots, X_n$ in non-decreasing order. It is well-known that the CDF of $X_{(j)}$ is $F_{X_{(j)}}(y) = G(y, j, n - j + 1)$, where $G(x, a, b)$ is the CDF of the $\text{Beta}(a, b)$ distribution evaluated at x, also known as the regularized incomplete beta function, $I_x(a, b)$. Furthermore, for any distinct $i, j \in \{1, \ldots, n\}$, the CDF of the conditional distribution of $X_{(j)}$ given $X_{(i)}$ is

$$F_{X_{(j)}|X_{(i)}=x}(y) = \begin{cases} G((y - x)/(1 - x), j - i, n - j + 1) & \text{if } i < j, \\ G(y/x, j, i - j) & \text{if } i > j; \end{cases} \tag{1}$$

see Theorems 2.4.1 and 2.4.2 of Arnold et al. (2008) for reference. We apply the probability integral transform to these marginal and conditional distributions. Specifically, defining

$$h_{ij}(x, y) = \begin{cases} F_{X_{(j)}}(y) & \text{if } i = j, \\ F_{X_{(j)}|X_{(i)}=x}(y) & \text{if } i \neq j, \end{cases}$$

and letting $U_{ij} := h_{ij}(X_{(i)}, X_{(j)})$, it follows that $U_{ij} \sim \text{Uniform}(0, 1)$. Values of U_{ij} close to 0 or 1 indicate that $X_{(j)}$ is smaller or larger than expected, given $X_{(i)}$ when $i \neq j$ or marginally when $i = j$. Thus, we construct a p-value $p_{ij} := 2 \min(U_{ij}, 1 - U_{ij})$, which is small whenever $X_{(j)}$ is unusually small or large, relative to $X_{(i)}$.

Now, suppose $\mathcal{I} = ((i_1, j_1), \ldots, (i_m, j_m))$ is a sequence of m pairs (i, j). We propose to compute the p-values p_{ij} as defined above for all (i, j) in $\mathcal{I}$ and then aggregate them using the Cauchy combination technique (Liu & Xie, 2020) to obtain a combined p-value,

$$p := 1 - F_{\text{Cauchy}}\left(\frac{1}{m} \sum_{(i,j) \in \mathcal{I}} F^{-1}_{\text{Cauchy}}(1 - p_{ij})\right) \tag{2}$$

where F_{Cauchy} is the CDF of the $\text{Cauchy}(0, 1)$ distribution. Finally, to adjust for the fact that the Cauchy combination does not exactly control Type I error due to the dependence among the p_{ij} values, we use a simple multiplicative correction, $p^* := \min(1, 1.15p)$. Empirically, we find that p^* approximately controls Type I error at any level $\alpha \in (0, 1)$, while attaining Type I error rates close to α when $\alpha \leq 0.05$; see Figure S3.

In Section S4, we provide a step-by-step algorithm for computing the PITOS p-value p^*.

2.2. *Choosing a sequence of pairs*

The PITOS test requires specification of a sequence of pairs $\mathcal{I}$, and the choice of $\mathcal{I}$ affects the power to detect different alternative hypotheses. Based on simulation studies, we find that any given pair (i, j) may be powerful for detecting certain alternatives, but not others. Furthermore, the power of (i, j) is not equal to the power of (j, i), in general. Roughly speaking, the power of a pair (i, j) is higher for departures from the null between the i/n and j/n quantiles. In other words, (i, j) tends to have good power to detect an alternative with density higher or lower than 1 in the interval from i/n to j/n. In particular, values of j/n close to 0 or 1 tend to be useful for detecting departures in the left or right tail, respectively.

Thus, to obtain good power across a wide range of alternatives, we aim to choose a sequence of points (i, j) such that $(i/n, j/n)$ is well-distributed throughout the unit square $[0, 1]^2$, with heavier coverage near the boundaries, since detecting tail deviations is of particular importance in many practical applications. It is also desirable to include additional points (i, i) on the diagonal, since the marginal order statistics tend to have good power for distributions close to Uniform$(0, 1)$. Note that a given point (i, j) may appear more than once in $\mathcal{I}$, and this has the effect of upweighting the contribution of that pair in the Cauchy combination of p-values in Equation 2.

Based on these considerations, we propose to generate $\mathcal{I}$ as follows. For $k = 1, \ldots, m - n$, generate (u_k, v_k) from the two-dimensional Halton sequence on $[0, 1]^2$ with bases 2 and 3 (Halton, 1964). Transform to $x_k = F^{-1}_{\text{Beta}(0.7,0.7)}(u_k)$ and $y_k = F^{-1}_{\text{Beta}(0.7,0.7)}(v_k)$, that is, map u_k and v_k through the inverse CDF of the Beta$(0.7, 0.7)$ distribution, and set $i_k = \lceil nx_k \rceil$ and $j_k = \lceil ny_k \rceil$. Finally, set $i_{m-n+r} = r$ and $j_{m-n+r} = r$ for $r = 1, \ldots, n$, and define $\mathcal{I} = ((i_1, j_1), \ldots, (i_m, j_m))$. To keep the algorithm as computationally efficient as possible, it makes sense to choose $m = O(n \log n)$ since computing the order statistics $X_{(1)}, \ldots, X_{(n)}$ already takes $O(n \log n)$ time. We recommend using $m = \lceil 10\, n \log n \rceil + n$ since we find that this provides good empirical performance. See Section S4 for a step-by-step algorithm.

Alternatively, one could use random samples of $(u_k, v_k) \sim \text{Uniform}([0, 1]^2)$, but the Halton sequence has the advantage of producing points that are more evenly distributed than a random sequence, and has the added benefit of being deterministic. The procedure above could also be customized by replacing the Beta$(0.7, 0.7)$ CDF with any other CDF on $[0, 1]$, in order to try to increase power for a specific collection of alternative distributions of interest.

3. Experiments

In this section, we compare the performance of PITOS (our method) and several benchmark GoF tests in simulation studies. As benchmarks, we compare with the Anderson–Darling (AD), Neyman–Barton (NB), Kolmogorov–Smirnov (KS), and Cramér–von Mises (CvM) tests; see Section S1 for details. To quantify the highest power possible, we also compare with an oracle benchmark based on the likelihood ratio test (LRT) using the true distribution as a point alternative; see Section S1. This represents the optimal power since the LRT is the most powerful test for a specific null/alternative pair, by the Neyman–Pearson lemma (Neyman & Pearson, 1933), but this test cannot be used in practice because it requires knowing the true distribution to compute the test statistic.

First, we consider the power of each test on the distributions from Figure 1, which represent motivating examples where the commonly used tests generally have low power. For a given sample size n, we compute the power to reject the null at the $\alpha = 0.05$ level by generating 100,000 simulated i.i.d. data sets from the true distribution and calculating the proportion of times that the p-value is less than $\alpha = 0.05$. Figure 2 shows the power of each test as a function of

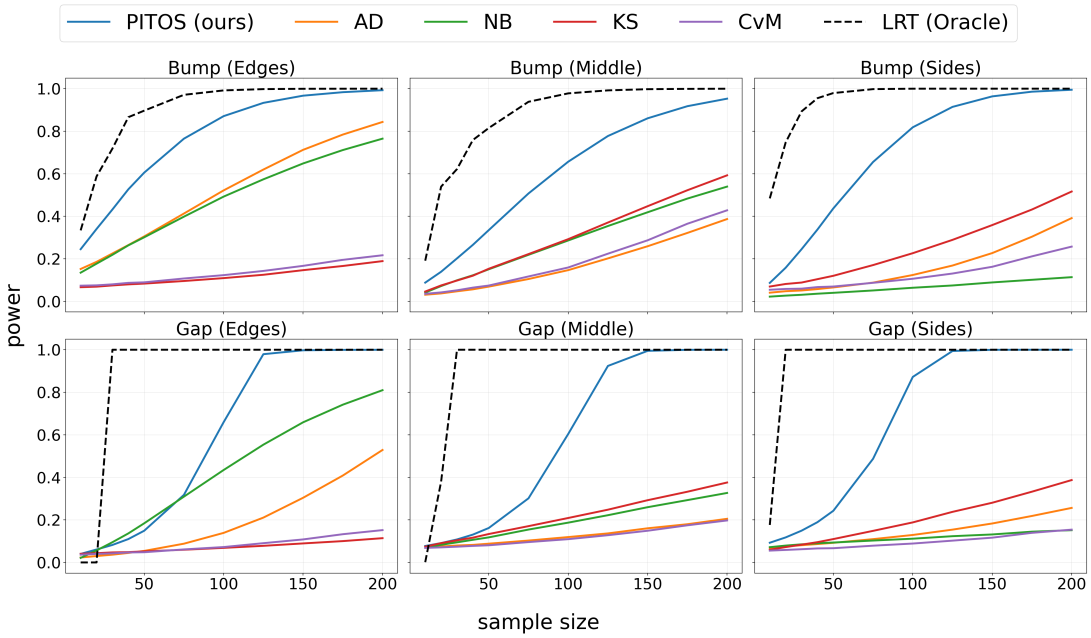

Fig. 2: Power of each method versus sample size n for the distributions in Figure 1.

n. We see that PITOS outperforms the other tests by a wide margin on all of these distributions. The value of these curves at $n = 200$ corresponds to the values of the power shown in Table 1.

Next, we consider the set of six distributions shown in Figure 3 (top), chosen to represent a range of behaviors. Figure 3 (bottom) shows the power at level $\alpha = 0.05$ versus n for each distribution, again using 100,000 simulated i.i.d. data sets for each n. On Uniform$(0, 1)$ data—that is, when the null hypothesis is true—all methods have power close to 0.05, indicating that they correctly control Type I error. (The LRT power is 0 here because the test statistic is identically constant.) The Beta$(1.2, 0.8)$ distribution is moderately close to uniform, and here we see that PITOS has the lowest power, although the differences between methods are relatively small. Meanwhile, PITOS is best or second-best on Beta$(0.6, 0.6)$ and Beta$(1.6, 1.6)$, which represent cases with somewhat higher or lower density, respectively, in the tails. PITOS excels when there are outliers or heavy tails, such as in the case of $\Phi($Laplace$(0, 1))$, which is the distribution of $\Phi(Y)$ where $Y \sim$ Laplace$(0, 1)$ and $\Phi(\cdot)$ is the standard normal CDF. Furthermore, PITOS tends to have high power to detect discreteness, such as in the case of Uniform$(\{0.01, 0.02, \ldots, 0.99\})$, the discrete uniform distribution on $\{0.01, 0.02, \ldots, 0.99\}$. The other methods have power essentially at $\alpha = 0.05$ on this discrete example, meaning that they completely fail to detect this departure from the null of Uniform$(0, 1)$.

Finally, we consider eight scenarios in which we randomly select a distribution P_θ from a given parametric family, and then approximate the power of each test on data from P_θ by simulating multiple i.i.d. data sets from P_θ. The scenarios considered are (1) Symmetric Heavy-Tailed, (2) Symmetric Light-tailed, (3) Asymmetric Heavy-tailed, (4) Asymmetric Light-tailed, (5) Outliers, (6) Nearly Uniform, (7) Random Bump, and (8) Random Gap; see Section S3 for details. For each scenario, we randomly select 1,000 distributions and, for each distribution, we approximate the power of each test by simulating 1,000 i.i.d. data sets of size $n = 100$. We summarize the results by aggregating in two ways. First, for each P_θ, we rank the tests according to their power, and

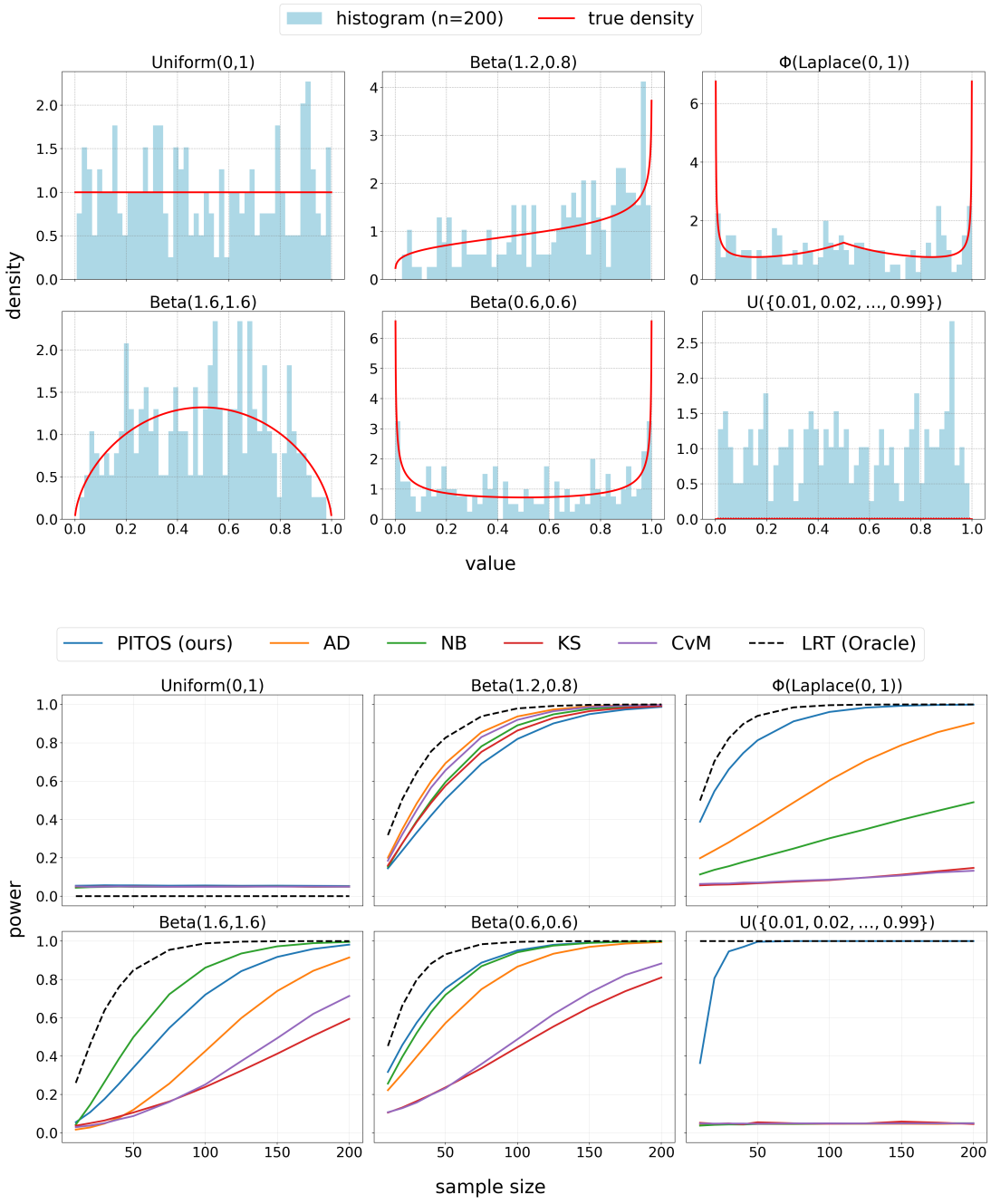

Fig. 3: Power of GoF tests across various data distributions.

compute the proportion of times that test t has rank r across distributions P_θ. Second, we simply compute the average power of each test across distributions P_θ.

Figure 4 shows the results. We see a similar pattern of performance as in Figures 2 and 3. In the Nearly Uniform scenario, PITOS performs worse than AD, NB, and CvM; however, all of the tests exhibit average power between 0.28 and 0.34, so the difference in performance is relatively

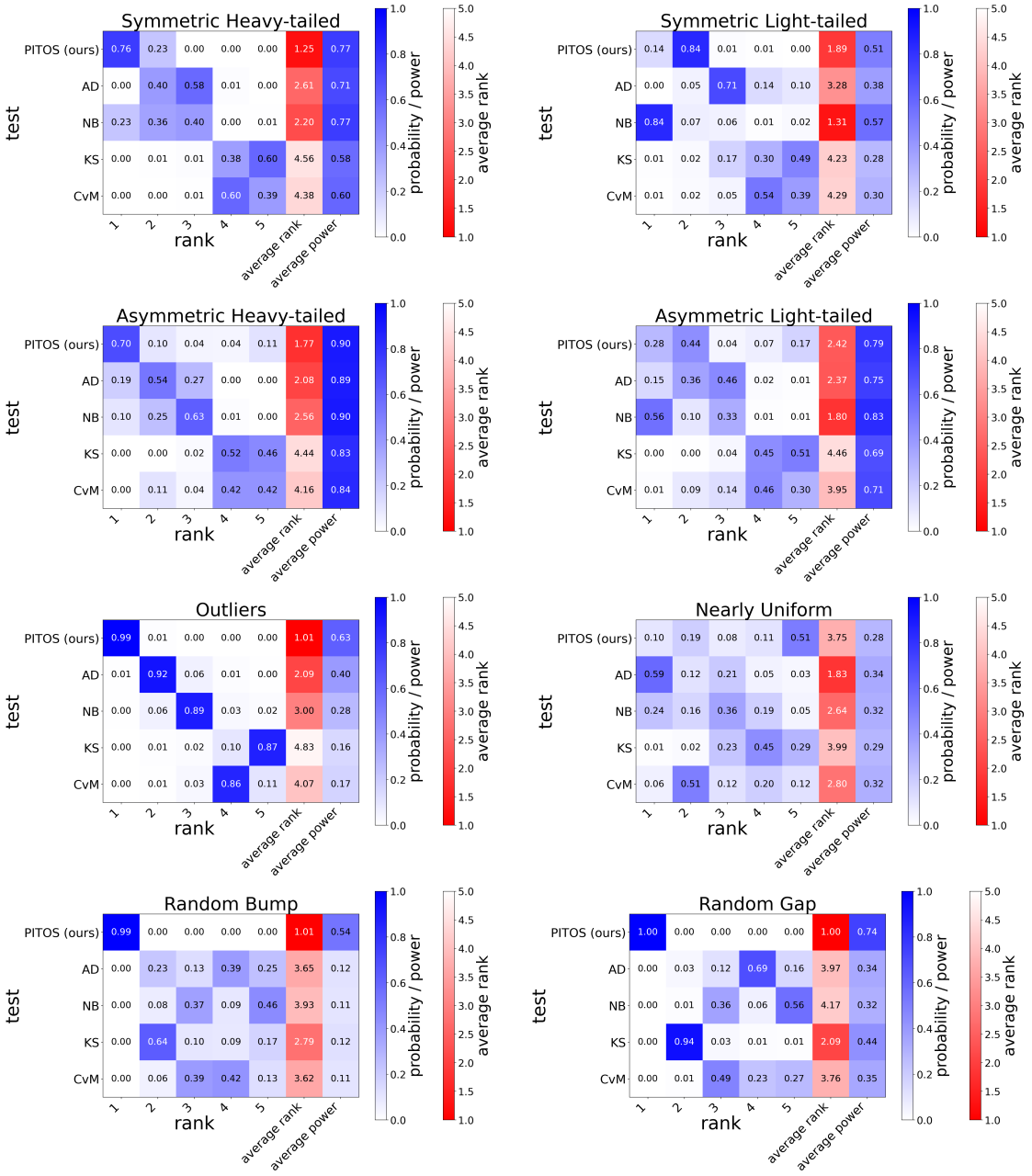

Fig. 4: Power of GoF tests across families of data distributions.

small. PITOS is usually outperformed by NB in the light-tailed scenarios, but it typically ranks first or second in these scenarios. In all of the other scenarios, PITOS dominates the rankings, particularly in the Outliers, Random Bump, and Random Gap scenarios.

4. Discussion

We have introduced the PITOS test, and demonstrated that it is competitive with existing GoF tests across a wide range of data distributions. Moreover, for distributions characterized by large local deviations from uniformity, PITOS is often substantially more powerful than existing methods. While we recommend running PITOS using our proposed sequence of pairs based on a transformed Halton sequence, the test can be customized by using different sequences in order to target certain types of alternative distributions. Optimizing the choice of sequence to maximize power for a given set of alternative distributions is a potentially interesting direction for future work. A related direction for future work would be to establish theoretical results on the power of the PITOS test.

Acknowledgements

C.T.C. was supported by National Institutes of Health (NIH) Training Grant T32CA09337. J.W.M. was supported in part by the National Cancer Institute of the NIH under award number R01CA240299. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Anderson, T. W. & Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* , 193–212.

Arnold, B. C., Balakrishnan, N. & Nagaraja, H. N. (2008). *A First Course in Order Statistics*. SIAM.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review* **59**, 65–98.

Brockwell, A. (2007). Universal residuals: A multivariate transformation. *Statistics & Probability Letters* **77**, 1473–1478.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal* **1928**, 13–74.

D'Agostino, R. B. & Stephens, M. A. (1986). *Goodness-of-Fit Techniques*. USA: Marcel Dekker, Inc.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM* **7**, 701–702.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att* **4**, 89–91.

Kornblith, S. et al. (2018). HypothesisTests.jl: Hypothesis tests for Julia. `https://github.com/JuliaStats/HypothesisTests.jl`. Version 0.11.6.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.

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

Neyman, J. (1937). "Smooth test" for goodness of fit. *Scandinavian Actuarial Journal* **1937**, 149–199.

Neyman, J. & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**, 289–337.

Von Mises, R. (1936). *Wahrscheinlichkeit Statistik und Wahrheit: Einführung in die neue Wahrscheinlichkeitslehre und ihre Anwendung*. Springer.

Supplementary Material for "A powerful goodness-of-fit test using the probability integral transform of order statistics"

S1. Standard Goodness-of-Fit Tests

In Table S1, we summarize the commonly used goodness-of-fit (GoF) tests that serve as benchmarks in our comparisons. For each of these tests, the data $X_1, \ldots, X_n$ are assumed to be i.i.d. and the null hypothesis is that $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. In the Neyman–Barton test, $\pi_1(x)$ and $\pi_2(x)$ are the first two orthogonal Legendre polynomials on $[0, 1]$.

Table S1: Common GoF Tests for Uniform$(0, 1)$ null hypothesis.

Test	**Test statistic** $T(X_1, \ldots, X_n)$
Anderson–Darling (AD) (Anderson & Darling, 1952)	$-n - n^{-1} \sum_{i=1}^{n} (2i-1)(\log X_{(i)} + \log(1 - X_{(n-i+1)}))$
2^{nd}-order Neyman–Barton (NB) (Neyman, 1937)	$\sum_{j=1}^{2} (n^{-1/2} \sum_{i=1}^{n} \pi_j(X_i))^2$ where $\pi_1(x) = 2\sqrt{3}x$ and $\pi_2(x) = \sqrt{5}(6x^2 - 0.5)$
Kolmogorov–Smirnov (KS) (Kolmogorov, 1933)	$\sup_{t \in [0,1]} \lvert F_n(t) - t \rvert$ where $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le t)$
Cramér–von Mises (CvM) (Cramér, 1928; Von Mises, 1936)	$1/(12n) + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - X_{(i)}\right)^2$
Likelihood ratio test (LRT) (Neyman & Pearson, 1933)	$\sum_{i=1}^{n} \log f_1(X_i)$ where $f_1(x)$ is the density of the alternative distribution

For the Neyman-Barton, Cramér–von Mises, and likelihood ratio tests, we implement the test statistic in the Julia programming language (Bezanson et al., 2017) and compute p-values using empirical null distributions computed based on 100,000 simulated i.i.d. data sets from the Uniform$(0, 1)$ distribution, for any given n. For Anderson–Darling and Kolmogorov–Smirnov, we use existing implementations from HypothesisTests.jl library in Julia (Kornblith et al., 2018).

S2. Generalized Rosenblatt Transform

For a random vector $Y_{1:n} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, define the conditional CDFs $F_1, \ldots, F_n$ as $F_1(y_1) = P(Y_1 \le y_1)$ and $F_k(y_k \mid y_{1:(k-1)}) = P(Y_k \le y_k \mid Y_{1:(k-1)} = y_{1:(k-1)})$ for $k \in \{2, 3, \ldots, n\}$. Likewise, define $F_1^-(y_1) = P(Y_1 < y_1)$ and $F_k^-(y_k \mid y_{1:(k-1)}) = P(Y_k < y_k \mid Y_{1:(k-1)} = y_{1:(k-1)})$ for $k \in \{2, 3, \ldots, n\}$.

For $y_1, \ldots, y_n \in \mathbb{R}$ and $u_1, \ldots, u_n \in (0, 1)$, define the function $h(y_{1:n}, u_{1:n}) \in \mathbb{R}^n$ such that

$$
\begin{aligned}
h_1(y_{1:n}, u_{1:n}) &= u_1 F_1(y_1) + (1 - u_1) F_1^-(y_1) \\
h_k(y_{1:n}, u_{1:n}) &= u_k F_k(y_k \mid y_{1:(k-1)}) + (1 - u_k) F_k^-(y_k \mid y_{1:(k-1)})
\end{aligned}
$$

for $k \in \{2, 3, \ldots, n\}$. If $U_1, \ldots, U_n$ i.i.d. $\sim \text{Uniform}(0, 1)$ and $X_{1:n} = h(Y_{1:n}, U_{1:n})$ then $X_1, \ldots, X_n$ are i.i.d. Uniform$(0, 1)$ random variables. A proof is provided by Brockwell (2007).

S3. Randomly Generated Distributions in Simulation Study

Table S2 describes the sampling process used to generate randomized distributions for each scenario in Figure 4. Here, δ_x denotes the unit point mass at x, $U(a,b)$ denotes the uniform distribution on the interval (a,b), and Gamma(α,β) denotes the Gamma distribution with density $f(x) = x^{\alpha-1}\exp(-x/\beta)/(\beta^\alpha\Gamma(\alpha))$ for $x > 0$. Visualizations of several randomly sampled densities for each scenario are shown in Figure S1.

Table S2: Procedures for randomly sampling distributions for each scenario in Figure 4.

Scenario	Parametric family	Random generation of parameters
Symmetric Heavy-tailed	$\text{Beta}(\mu\sigma, (1-\mu)\sigma)$	$\mu \sim \delta_{1/2}$, $\sigma \sim \text{Gamma}(3, 1/2)$ Draw (μ,σ) until $\min(\mu\sigma,(1-\mu)\sigma) \le 1$
Symmetric Light-tailed	$\text{Beta}(\mu\sigma, (1-\mu)\sigma)$	$\mu \sim \delta_{1/2}$, $\sigma \sim \text{Gamma}(5, 1/2)$ Draw (μ,σ) until $\min(\mu\sigma,(1-\mu)\sigma) > 1$
Asymmetric Heavy-tailed	$\text{Beta}(\mu\sigma, (1-\mu)\sigma)$	$\mu \sim \text{Beta}(2,2)$, $\sigma \sim \text{Gamma}(3, 1/2)$ Draw (μ,σ) until $\min(\mu\sigma,(1-\mu)\sigma) \le 1$
Asymmetric Light-tailed	$\text{Beta}(\mu\sigma, (1-\mu)\sigma)$	$\mu \sim \text{Beta}(2,2)$, $\sigma \sim \text{Gamma}(5, 1/2)$ Draw (μ,σ) until $\min(\mu\sigma,(1-\mu)\sigma) > 1$
Outliers	$\pi\, U(0,b) + (1-\pi)\, U(0,1)$	$\pi \sim U(0, 0.1)$, $b \sim U(0, 0.01)$
Nearly Uniform	$\text{Beta}(\mu\sigma, (1-\mu)\sigma)$	$\mu \sim \text{Beta}(50,50)$, $\sigma \sim \text{Gamma}(100, 1/50)$
Random Bump	$\pi\, U(m-0.001, m+0.001) + (1-\pi)\, U(0,1)$	$m \sim U(0.001, 0.999)$, $\pi \sim U(0, 0.1)$
Random Gap	$\left(\frac{m-w}{1-2w}\right) U(0, m-w) + \left(\frac{1-(m+w)}{1-2w}\right) U(m+w, 1)$	$m \sim U(0.1, 0.9)$, $w \sim U(0.025, 0.1)$

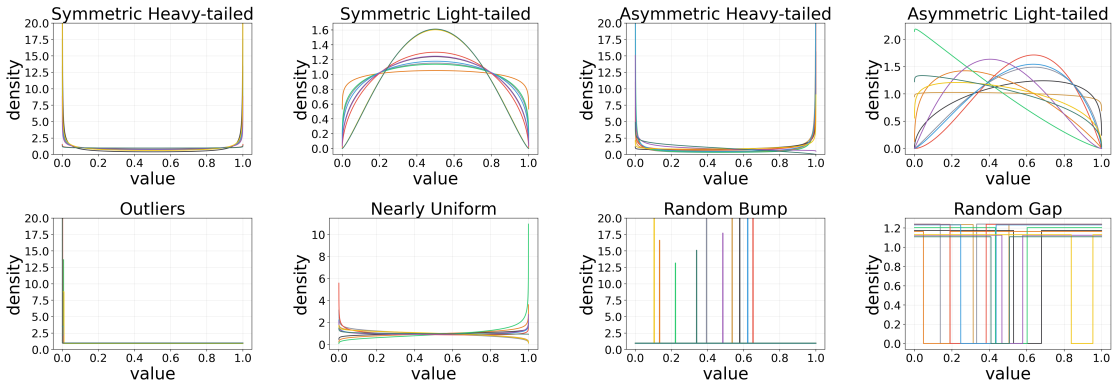

Fig. S1: Examples of randomly sampled densities for each scenario in Figure 4.

S4. Algorithms for Implementing the PITOS Goodness-of-Fit Test

In this section, we provide step-by-step algorithms for computing the PITOS p-value p^* (Algorithm S1), generating Halton sequences (Algorithm S2), and our proposed method of generating the sequence of pairs to be used in the PITOS algorithm (Algorithm S3).

S4.1. *Main PITOS algorithm*

The algorithm takes $O(n \log n)$ time for any collection of pairs $\mathcal{I}$ of size $O(n \log n)$. We write $G(x, a, b)$ to denote the CDF of Beta(a, b) at x, that is, $G(x, a, b) = \int_0^x \frac{1}{\mathrm{B}(a,b)} t^{a-1}(1-t)^{b-1} dt$.

Algorithm S1 PITOS goodness-of-fit test for Uniform$(0, 1)$ null distribution

Input: Data $x_1, \ldots, x_n \in [0, 1]$ and pairs $\mathcal{I} = ((i_1, j_1), \ldots, (i_m, j_m))$ where $i_k, j_k \in \{1, \ldots, n\}$
Output: Approximate p-value p^*

1. Sort $x_1, \ldots, x_n$ to obtain the order statistics $x_{(1)} \le \cdots \le x_{(n)}$
2. **for** $(i, j) \in \mathcal{I}$ **do**
   - **if** $i = j$ **then**
     - $u_{ij} \leftarrow G(x_{(j)}, j, n - j + 1)$
   - **elseif** $i < j$ **then**
     - $u_{ij} \leftarrow G((x_{(j)} - x_{(i)})/(1 - x_{(i)}), j - i, n - j + 1)$
   - **else**
     - $u_{ij} \leftarrow G(x_{(j)}/x_{(i)}, j, i - j)$
   - **end**
   - $p_{ij} \leftarrow 2 \min(u_{ij}, 1 - u_{ij})$

   end
3. $p \leftarrow 1 - F_{\text{Cauchy}}\left(\frac{1}{m} \sum_{(i,j)\in\mathcal{I}} F^{-1}_{\text{Cauchy}}(1 - p_{ij})\right)$
4. $p^* \leftarrow \min(1, 1.15p)$
5. **return** p^*

Since a given pair (i, j) may occur more than once in $\mathcal{I}$, this algorithm could potentially be sped up further by caching the value of p_{ij} for each (i, j) as it is computed, and re-using the previously computed value of p_{ij} instead of re-computing it if (i, j) is encountered again.

S4.2. *Generating PITOS pairs using a discretized transformed Halton sequence*

For completeness, we next provide a procedure for generating Halton sequences (Halton, 1964) in Algorithm S2, which is used as a subroutine in Algorithm S3. In Algorithm S2, we write $i \bmod b$ to denote the remainder when dividing i by b, where i and b are positive integers.

Algorithm S3 details how we compute the sequence $\mathcal{I}$ to be used in Algorithm S1. To visualize the sequence of pairs $\mathcal{I} = ((i_1, j_1), \ldots, (i_m, j_m))$ produced by Algorithm S3, we plot heatmaps in Figure S2 showing the number of times each (i, j) is included in $\mathcal{I}$, for a range of sample sizes $n \in \{25, 50, 100, 200\}$.

Algorithm S2 `Halton`: Generate one-dimensional element of Halton sequence

Input: Index $i \in \{1, 2, \ldots\}$ and base $b \in \{1, 2, \ldots\}$
Output: The ith element of the one-dimensional Halton sequence with base b

1. $f \leftarrow 1$
2. $x \leftarrow 0$
3. **while** $i > 0$ **do**
   $f \leftarrow f/b$
   $x \leftarrow x + (i \bmod b) \cdot f$
   $i \leftarrow \lfloor i/b \rfloor$
   end
4. **return** x

Algorithm S3 `GeneratePairs`: Generate sequence of pairs for PITOS test in Algorithm S1

Input: Sample size $n \in \{1, 2, \ldots\}$
Output: Sequence of pairs $\mathcal{I} = ((i_1, j_1), \ldots, (i_m, j_m))$ for input to PITOS test

1. $m \leftarrow \lceil 10\, n \log n \rceil + n$
2. **for** $k = 1, \ldots, m - n$ **do**
   $u \leftarrow \texttt{Halton}(k, 2)$
   $v \leftarrow \texttt{Halton}(k, 3)$
   $x \leftarrow F^{-1}_{\text{Beta}(0.7,0.7)}(u)$
   $y \leftarrow F^{-1}_{\text{Beta}(0.7,0.7)}(v)$
   $i_k \leftarrow \lceil nx \rceil$
   $j_k \leftarrow \lceil ny \rceil$
   end
3. **for** $r = 1, \ldots, n$ **do**
   $i_{m-n+r} \leftarrow r$
   $j_{m-n+r} \leftarrow r$
   end
4. **return** $((i_1, j_1), \ldots, (i_m, j_m))$

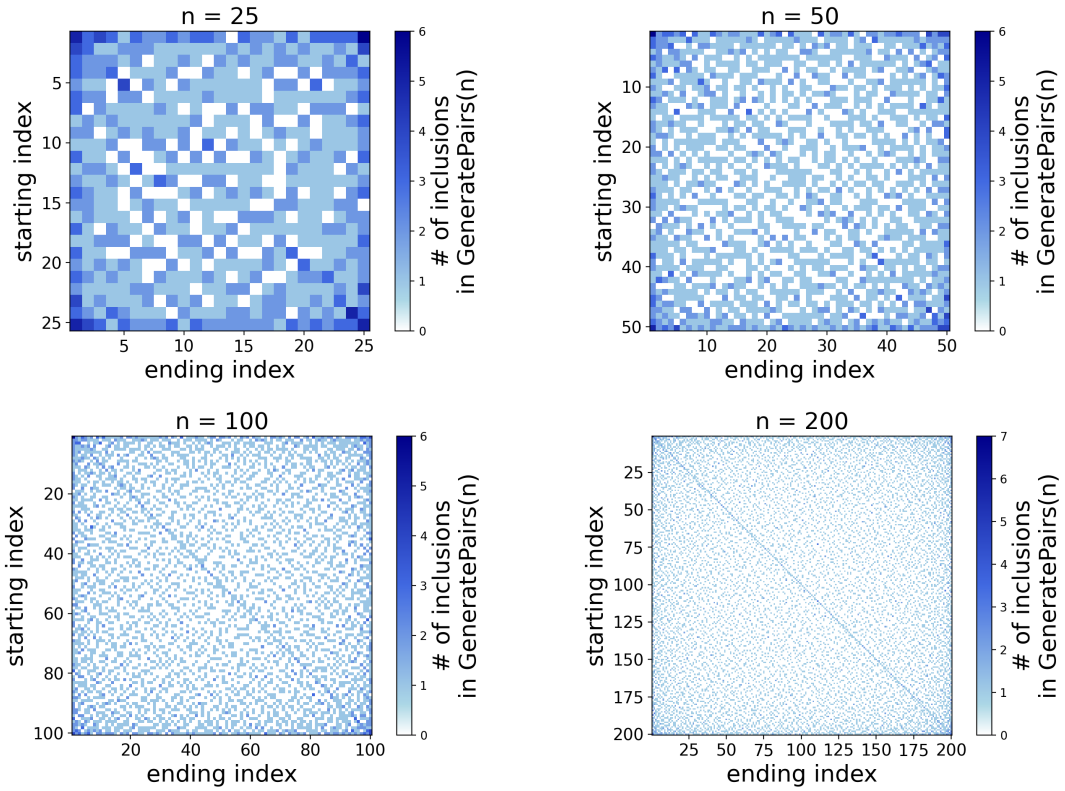

Fig. S2: Number of times each pair of indices (i, j) is included in `GeneratePairs(n)`.

S5. Approximately correct control of Type I error

Figure S3 shows the CDF of PITOS p-values with and without the 1.15× correction (that is, p^* and p, respectively), for i.i.d. data generated under the null hypothesis of $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$ with $n = 30$. A p-value CDF above the Uniform$(0, 1)$ CDF (shown as a black dashed line) is too liberal, meaning that Type I error is larger than α when rejecting the null at level α. Meanwhile, a p-value CDF below the Uniform$(0, 1)$ CDF is conservative, meaning that Type I error is smaller than α when rejecting at level α.

While the CDFs of both p and p^* are overly conservative for values greater than 0.2, they are fairly close to the Uniform$(0, 1)$ CDF for values between 0 and 0.1, which is the range of values that matters in practice. In the range $[0, 0.01]$, the CDF of the uncorrected p-values p is slightly liberal. Meanwhile, the CDF of the corrected p-values p^* is very close to the Uniform$(0, 1)$ CDF between 0 and 0.05, and is slightly conservative between 0.05 and 0.1.

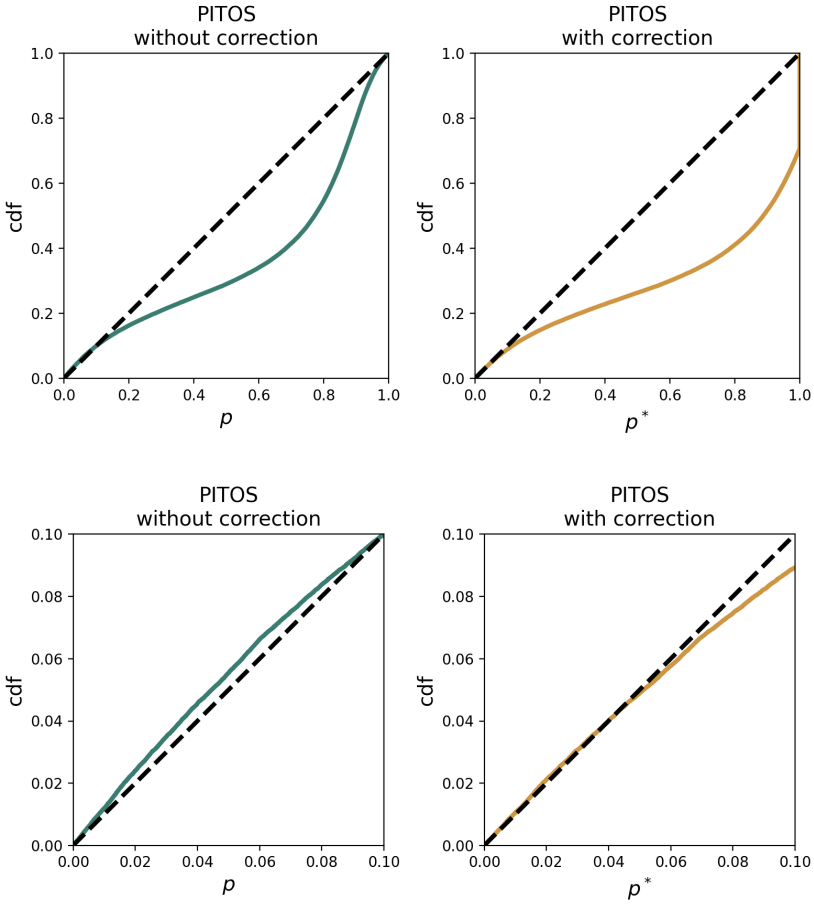

Fig. S3: CDFs of PITOS p-values with and without the 1.15× correction, under the null hypothesis, for a sample size of $n = 30$. Top plots show CDF over the full support of $[0, 1]$. Bottom plots are zoomed in to $[0, 0.1]$; in practice, it is common to reject the null at levels $\alpha \in (0, 0.05]$.