

## THE RÉNYI OUTLIER TEST

RYAN CHRIST, IRA HALL, AND DAVID STEINSALTZ

ABSTRACT. Cox and Kartsonaki proposed a simple outlier test for a vector of  $p$ -values based on the *Rényi transformation* that is fast for large  $p$  and numerically stable for very small  $p$ -values – key properties for large data analysis. We propose and implement a generalization of this procedure we call the Rényi Outlier Test (ROT). This procedure maintains the key properties of the original but is much more robust to uncertainty in the number of outliers expected *a priori* among the  $p$ -values. The ROT can also account for two types of prior information that are common in modern data analysis. The first is the prior probability that a given  $p$ -value may be outlying. The second is an estimate of how far of an outlier a  $p$ -value might be, conditional on it being an outlier; in other words, an estimate of effect size. Using a series of pre-calculated spline functions, we provide a fast and numerically stable implementation of the ROT in our R package `renyi`.

Cox and Kartsonaki proposed an outlier test based on the *Rényi transformation*  $\rho : [0, 1]^p \rightarrow [0, \infty]^p$  Cox and Kartsonaki (2019). For an ordered vector  $u \in [0, 1]^p$ , such that  $u_1 \leq u_2 \leq \dots \leq u_p$ ,  $\rho(u)_j = j \log(u_{j+1}/u_j)$  for all  $j < p$  and  $\rho(u)_p = -p \log(u_p)$ . Alfréd Rényi pointed out that when  $\rho$  is applied to a vector  $U$  of ordered independent uniform random variables, the image  $\rho(U)$  has entries that are independent exponential random variables Rényi (1953). Building on this observation, for some user-specified number of potential outliers,  $k$ , Cox and Kartsonaki proposed testing the null hypothesis  $H_0$  that an observed  $u$  is a vector of ordered independent uniform random variables by comparing  $\sum_{j=1}^k \rho(u)_j$  against its null distribution: a Gamma distribution with shape  $k$  and rate 1. This simple procedure allows for the rapid calculation of numerically precise  $p$ -values even when  $p$  is very large and the  $p$ -value returned is in the lower ranges accessible to machine precision. However, the power depends sensitively on the *a priori* specification of the number of outliers  $k$ . A more robust method would maintain power in the more common situation where the number  $k$  of outliers is unknown, but it is possible to specify a rough upper bound  $K$  on the likely number of outliers.

We present a robust generalization of Cox and Kartsonaki’s proposal that only requires an approximate upper bound  $K$ . Our generalization also admits two types of prior information that is common in modern applications can be used to sharpen the alternative hypothesis and thereby improve power. The first,  $\pi \in \mathbb{R}_{\geq 0}^p$ , is taken to be proportional to the prior probability that a given uniform random variable is an outlier. The second,  $\eta \in \mathbb{R}_{\geq 0}^p$ , is related to effect size: how far outlying  $u_j$  will be given that it is an outlier. In the common context where each element of  $u$  can be thought of as a  $p$ -value for testing whether some coefficient  $\beta$  in a linear regression model is zero, we take  $\eta_j \propto \mathbb{E}[\beta_j^2 | \beta_j \neq 0]$ . In the absence of prior

---

*Key words and phrases.* outlier test,  $p$ -value combination, Higher Criticism, sparse, global null.

information or expectations, we take the neutral defaults  $\pi_j = 1$  and  $\eta_j = 1$  for all  $j$ . Critically, our approach, which we call the Rényi Outlier Test (ROT), maintains the computational speed and numerical precision of the original test proposed by Cox and Kartsonaki. We also provide the `renyi` R package that implements our procedure, making use of pre-calculated spline functions. The package is publicly available at [ryanchrist.r-universe.dev/renyi](https://ryanchrist.r-universe.dev/renyi).

Compared to more commonly used “minimum”-based approaches, such as testing the minimum p-value with Bonferroni correction or Holm’s method, Higher Criticism and related tests in the General Goodness of Fit test family have more power when, roughly speaking, there are a handful of modestly small p-values Donoho and Jin (2004); Zhang et al. (2020); Zhang and Wu (2022). Computational speed and numerical precision have been major obstacles to applying these outlier tests in practice. Recently, Wang et al. proposed a fast implementation of higher criticism that is numerically stable for even very small p-values Wang et al. (2024). However, this approach does not admit prior information such as  $\pi$  and  $\eta$ .

Given an initial set of *unordered* uniform random variables  $U \in [0, 1]^p$ ,  $\pi$ , and  $\eta$ , the ROT is a two step procedure to test the null hypothesis  $H_0 : U_j \stackrel{iid}{\sim} \text{Unif}(0, 1)$  for  $j = 1, \dots, p$ . Note that if each  $U_j$  represents a p-value, they must be **exactly uniform** under the global null hypothesis, not sub-uniform or super-uniform. First, we perform a simple generalization of the *Rényi transformation* which accounts for  $\pi$  and  $\eta$  to obtain a set of independent standard exponential random variables. We then test the outliers based on those exponential random variables using a procedure robustified to our choice of  $K$ .

Define

$$(1) \quad Z_j = \eta_j (-\log(U_j) + \log(\pi_j)) = \eta_j (-\log(U_j)) + \zeta_j$$

where  $\zeta_j = \eta_j \log(\pi_j)$ , and let  $N : \mathbb{R} \rightarrow \mathbb{N}$  be the corresponding point process:

$$(2) \quad N(t) := \sum_{j=1}^p \mathbf{1}\{Z_j \leq t\}.$$

For  $-\infty \leq t < \infty$ , define a filtration by letting  $\mathcal{F}_t$  be the sigma algebra generated by all events of the form

$$\left\{ -\log(U_j) \leq \frac{s}{\eta_j} - \log \pi_j \right\}$$

for  $s \leq t$  and  $1 \leq j \leq p$ . We understand  $\eta_j$  and  $\pi_j$  to be measurable with respect to  $\mathcal{F}_t$  for all  $t$  (including  $t = -\infty$ ).  $N(t)$  is adapted with respect to this filtration, and the compensator is

$$(3) \quad \Lambda(t) := \sum_{j=1}^p \eta_j^{-1} (t \wedge Z_j - t \wedge \zeta_j).$$

Since  $\Lambda : \mathbb{R} \rightarrow [0, -\sum \log U_j]$  is a continuous non-decreasing function, it has a right-inverse  $\Lambda^{-1} : [0, -\sum \log U_j] \rightarrow [\min \zeta_j, \max Z_j]$  defined by  $\Lambda^{-1}(u) = \sup\{t : \Lambda(t) < u\}$ ; that is,  $\Lambda \circ \Lambda^{-1}$  is the identity map on  $[0, -\sum \log U_j]$ . Then by Theorem 15.15 of Kallenberg (2021)  $N \circ \Lambda^{-1}$  is a Poisson process with unit rate (up to the time of the  $p$ -th event). The test will then be based on the statistics  $(X_1, \dots, X_p)$ , which are the interarrival times of the process, in reverse order; under the global null hypothesis these are i.i.d. unit exponential random variables. This is a generalization of the original *Rényi transformation*.

Let  $G_x$  denote the CDF of the Gamma distribution with shape parameter  $x$  and rate 1, and let  $I_{x,y}$  denote the CDF of the Beta distribution with mean  $x/(x+y)$ .

If the number of outliers  $k$  were known, then the p-value  $1 - G_k \left( \sum_{j=1}^k X_j \right)$  would provide a well-powered test of  $H_0$  against alternatives where  $k$  of the original values  $U_j$  are sampled from a distribution that makes them substantially smaller than uniform  $[0,1]$  random variables. If  $k$  is chosen too small then some potential power is lost, while a too-large  $k$  would crush the power by mixing the true outliers with non-outlying observations. To mitigate this weakness, then, we try to find an upper bound  $K$  on  $k$ , and perform an omnibus test over different  $k$  up to  $K^* := 2^{\lceil \log_2 K \rceil}$ . Let  $\tilde{X}_j = X_j$  for all  $j < K^*$  and define

$$(4) \quad \tilde{X}_{K^*} = -\log \left( 1 - I_{p-K^*+1, K^*} \left( \exp \left( - \sum_{j=K^*}^p \frac{X_j}{j} \right) \right) \right).$$

Using Rényi's representation for the order statistics of independent exponential random variables, the sum in (4) is distributed (under the global null hypothesis) as the  $K^*$ -th largest out of  $p$  independent unit exponential random variables Rényi (1953). Exponentiating this as above yields the corresponding order statistic of Uniform random variables, which can be transformed by the Beta cdf to a new Uniform random variable, leaving us at last with  $\tilde{X}_{K^*}$  being another unit exponential random variable, independent of  $\tilde{X}_1, \dots, \tilde{X}_{K^*-1}$ .

We now define the ROT test statistic as

$$(5) \quad \rho_{K^*} = \max_{i \in \mathcal{I}_k} -\log \left( 1 - G_i \left( \sum_{j=1}^i \tilde{X}_j \right) \right)$$

where  $\mathcal{I}_k = (1, 2, 4, 8, \dots, K^*)$ . The fact that each  $\tilde{X}_j$  in (5) is an independent exponential, makes simulating the null distribution of  $\rho_{K^*}$  straightforward. We used Monte Carlo simulation to estimate the body of the null distribution of  $\rho_{K^*}$  for  $K^*$  taking values in  $(1, 2, 4, \dots, 128)$ . We used those null simulations to fit a line to the log-linear tail of each distribution and fit a cubic spline function to the body of the distribution. This yielded a compressed form of a lookup table for each test statistic that allows rapid computation p-values for a wide range of  $K^*$ . It is available via our R package `renyi`. The package is publicly available at `ryanchrist.r-universe.dev/renyi`.

## REFERENCES

- David R Cox and Christiana Kartsonaki. On the analysis of large numbers of p-values. *International Statistical Review*, 87(3):505–513, 2019.
- David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.
- Olaf Kallenberg. *Foundations of Modern Probability*. Springer Verlag, third edition, 2021.
- Alfréd Rényi. On the theory of order statistics. *Acta Math. Acad. Sci. Hung.*, 4(2), 1953.
- Wenjia Wang, Yusi Fang, Chung Chang, and George C Tseng. Accurate and ultra-efficient p-value calculation for higher criticism tests. *Journal of Computational and Graphical Statistics*, 33(2):463–476, 2024.
- Hong Zhang and Zheyang Wu. The general goodness-of-fit tests for correlated data. *Computational Statistics & Data Analysis*, 167:107379, 2022.
- Hong Zhang, Jiashun Jin, and Zheyang Wu. Distributions and power of optimal signal-detection statistics in finite case. *IEEE Transactions on Signal Processing*, 68:1021–1033, 2020.

(Ryan Christ and Ira Hall) CENTER FOR GENOMIC HEALTH & DEPARTMENT OF GENETICS, YALE UNIVERSITY SCHOOL OF MEDICINE, NEW HAVEN, CT USA

*Email address*, Ryan Christ: `ryan.christ@yale.edu`

*Email address*, Ira Hall: `ira.hall@yale.edu`

(David Steinsaltz) DEPARTMENT OF STATISTICS, OXFORD UNIVERSITY, OXFORD, UK

*Email address*, David Steinsaltz: `steinsal@stats.ox.ac.uk`