Published by IOP Publishing for Sissa Medialab

Received: *October 26, 2022*
Revised: *January 20, 2023*
Accepted: *February 27, 2023*
Published: *March 29, 2023*

A goodness-of-fit test based on a recursive product of spacings

Philipp Eller[a,b,*] **and Lolian Shtembari**[c,*]

[a] *Technical University Munich,*
Garching, Germany
[b] *Exzellenzcluster ORIGINS,*
Garching, Germany
[c] *Max Planck Institute for Physics,*
Munich, Germany

E-mail: philipp.eller@tum.de, lolian@mpp.mpg.de

Abstract: We introduce a new statistical test based on the observed spacings of ordered data. The statistic is sensitive to detect non-uniformity in random samples, or short-lived features in event time series. Under some conditions, this new test can outperform existing ones, such as the well known Kolmogorov-Smirnov or Anderson-Darling tests, in particular when the number of samples is small and differences occur over a small quantile of the null hypothesis distribution. A detailed description of the test statistic is provided including a detailed discussion of the parameterization of its distribution via asymptotic bootstrapping as well as a novel per-quantile error estimation of the empirical cumulative distribution. Two example applications are provided, using the test to boost the sensitivity in generic "bump hunting", and employing the test to detect supernovae. The article is rounded off with an extended performance comparison to other, established goodness-of-fit tests.

Keywords: Analysis and statistical methods; Data analysis; Pattern recognition, cluster finding, calibration and fitting methods; Trigger concepts and systems (hardware and software)

ArXiv ePrint: 2111.02252

[*] Corresponding author.

https://doi.org/10.1088/1748-0221/18/03/P03048

Contents

1 Introduction

Assessing the goodness-of-fit of a distribution given a number of random samples is an often-encountered problem in data analysis. Such statistical hypothesis tests find applications in many fields, ranging from the natural and social sciences over engineering to quality control. Several non-parametric tests exist, some of which have become standard tools, including the Kolgogorov-Smirnov (KS) test [1, 2] or the Anderson-Darling (AD) test [3]. [4] provides a comprehensive overview of existing tests, and a comparison of their performance for the case of detecting non-uniformity for a set of alternative distributions.

In this work, we are in contrast interested in the case where the bulk of samples are actually distributed according to the null hypothesis, and only few additional samples are introduced that are following a different distribution, representing a narrow excess over a known background. We present the new test statistic "recursive product of spacings", or short RPS, that is based on the spacings between ordered samples, and introduced in section 2. In section 3 we provide a parametrization of its distribution based on simulations, introducing techniques to estimate the asymptotic result of infinite bootstrapping steps in order to improve the quality of our fits. Subsequently we discuss the quality of the approximation deriving a per-quantile error estimate up to a desired confidence level.

The rest of the article focuses on some illustrations and example applications, as well as a detailed performance comparison to several other test statistics.

1.1 Goodness-of-fit tests

Suppose that we have obtained n samples y_i, and want to quantitatively test the hypothesis of those samples being random variates of a known distribution $f(y)$, i.e. independent and identically distributed (i.i.d.) according to $f(y)$. Here, we consider only continuous distributions $f(y)$ with cumulative $F(y)$, and hence can transform samples onto the unit interval $[0,1]$ via the Probability integral transformation $x_i = F(y_i)$ [5, 6]. This reduces the task at hand to test transformed samples x_i being distributed according to the standard uniform distribution $\mathcal{U}(0,1)$. Therefore, in the rest of this note, without loss of generality, we will only consider samples x_i assuming a uniform distribution as the null hypothesis.

First, let us briefly introduce other, existing test statistics to which we will compare the RPS statistic. We consider in particular two groups of statistics, those based on the empirical cumulative distribution (ECDF Statistics), and those based on the spacings between ordered samples (Spacings Statistics). An comprehensive overview of existing test statistics can be found in [4].

1.1.1 ECDF statistics

This class of test statistics compares the empirical cumulative distribution function (ECDF) $F_n(x)$ to the cumulative distribution function (CDF) $F(x)$, (here $F(x) = x$). Clustering of points under the null hypothesis of a uniform distribution would induce a steeper ECDF compared to the expected CDF, leading to a large deviation between the two. In particular, the following tests are widely used in order to detect such deviations:

- Kolmogorov-Smirnov (KS) [1, 2]: $D_n = \sup_x |F_n(x) - F(x)|$
- Cramer-von-Mises (CvM) [7, 8]: $T = n\int_{-\infty}^{\infty}(F_n(x) - F(x))^2 dF(x)$
- Anderson-Darling (AD) [3]: $A^2 = n\int_{-\infty}^{\infty}\frac{(F_n(x)-F(x))^2}{F(x)\,(1-F(x))}\,dF(x)$

Similar are a type of statistics defined on the ordered set. Given the n samples $\{x_1, x_2, \ldots, x_n\}$, we define the ordered set of samples as $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, where $x_{(i)} < x_{(i+1)}\ \forall i$. The expected value of ordered sample i is $i/(n+1)$, and we define the deviation to the expected values as $\delta_i = x_{(i)} - i/(n+1)$ for each sample i. Based on this we can write out the following two statistics:

- Pyke's Modified KS (C) [9, 10]: $C_n = \max(\max(\delta_i), -\min(\delta_i))$
- Brunk's Modified KS (K) [11]: $K_n = \max(\delta_i) - \min(\delta_i)$

1.1.2 Spacings statistics

Based on the ordered set, we can further define the $n+1$ spacings s as $s_i = x_{(i)} - x_{(i-1)}$, with $x_{(0)} = 0$ and $x_{(n+1)} = 1$. Clusters of points would lead to an increased number of unusually small spacings compared to the expectations, thus it is possible to construct tests sensitive to small spacings. Several such test statistics built from these spacings are considered in literature, including:

- Moran (M) [12]: $M = -\sum_{i=1}^{n+1}\log s_i$
- Greenwood (G) [13]: $G = \sum_{i=1}^{n+1} s_i^2$

The above two tests are defined as sums over functions of spacings, which allows to calculate the asymptotic distribution under the limit of large n using LeCam's theorem [14]. Other functions apart from the ones listed have been considered in literature, such as the difference of spacings compared to their expected values, or powers of this quantity. For a more extensive list of proposed tests, see [9]

In the context of a fixed rate Poisson process, these spacings can also be interpreted as *interarrival times* or *waiting times*. In some other areas, spacings are also referred to as *gaps*.

So-called *higher order spacings* can be defined by summing up neighbouring spacings. Here we consider the overlapping m-th order spacings $s_i^{(m)} = x_{(i+m)} - x_{(i)}$. With those, we can define generalisations of Moran and Greenwood, respectively, as discussed by Cressie:

- Logarithms of higher order spacings (Lm) [15]: $L_n^{(m)} = -\sum_{i=0}^{n-m+1} \log s_i^{(m)}$

- Squares of higher order spacings (Sm) [16]: $S_n^{(m)} = \sum_{i=0}^{n-m+1} (s_i^{(m)})^2$

For our comparisons presented later, we choose $m = 2$ and $m = 3$, respectively, to limit ourselves to a finite list of tests.

Other statistics based on spacings exist and are being actively developed and used, such as, for example, tests based on the k smallest or largest spacings [17].

2 Recursive product of spacings (RPS)

In this work, our goal is to construct a new test statistic, that has better sensitivity to narrow features or clusters in an otherwise uniform distribution of samples. The tell-tale sign we are looking for is a localized group of uncommonly small spacings of the ordered data. For this purpose, we propose a new class of test statistics, that are including higher order spacings in a recursive way.

The recursive product of spacings (RPS) can be thought of as an extension of the Moran statistic, and is defined as:

$$RPS(n) = \sum_{j=1}^{n} M_j = M_1 + M_2 + \cdots + M_n, \tag{2.1}$$

where the term M_1 is the *simple* sum of negative log spacings equivalent to the Moran statistic:

$$M_1 \equiv M = -\sum_{i=1}^{n+1} \log\left(s_{i,1}\right). \tag{2.2}$$

where $s_{i,1} \equiv s_i$ are the simple spacings considered before. The sum over all $\log(s_i)$ is the same as the logarithm of the product over all spacings s_i, thus the name *product* for the test. Additionally, working with logarithms is numerically more stable than products. All terms in eq. 2 are computed in the same way as Moran's test:

$$M_j = -\sum_{i=1}^{n+2-j} \log\left(s_{i,j}\right), \tag{2.3}$$

but with modified spacings $s_{i,j}$, defined for $1 < j \leq n$ as:

$$s^*_{i,j} = \frac{s_{i,j-1} + s_{i-1,j-1}}{2} \tag{2.4}$$

$$s_{i,j} = \frac{s^*_{i,j}}{\sum_i s^*_{i,j}} \tag{2.5}$$

which there are $n + 2 - j$ of, and that depend on the spacings $s_{i,j-1}$ used to compute the previous term M_{j-1} (hence the *recursiveness*). In order to better understand eq. (2.5) we can turn to figure 1, where we show how to transition from layer $j - 1$ (top) to layer j (bottom): in the top plot we show a list of events (blue), where we also highlight the boundaries 0 and 1 since they contribute to defining spacings; in the middle plot the middle points of the top row spacings are shown, forming a reduced set of "events", which is then transformed in order to ensure that the spacings of the new set sum up to 1, as shown in the bottom plot; the number of spacings going from the top plot to the bottom one is reduced by one, showing how we have a finite number of reduction steps in the definition of the RPS.

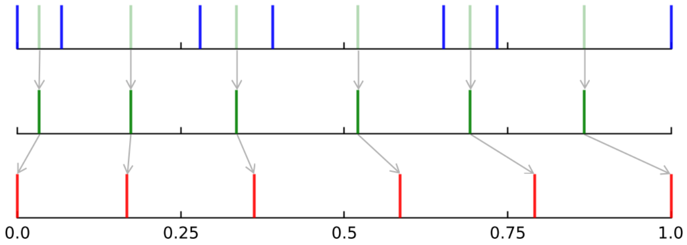


Figure 1. Example of the reduction step included in the RPS calculation. Given an initial set of events (top; blue), the middle points are calculated (top and middle; green) following eq. (2.4), which are then scaled in order to fill the $[0, 1]$ interval, forming a new set of data (bottom; red), following eq. (2.5). The evolution of sample positions on the $[0, 1]$ interval are annotated via the arrows.

Regarding eq. (2.3), we would like to point out its time reversal invariance: if the events were to be flipped ($\{x_i\} \to \{1 - x_i\}$), then one would obtain the same list of spacings in reversed order at all layers. The time reversal invariance in our formulas follows directly from the commutativity of sums and products.

We can see that term M_2 is identical to $L^{(2)}_n$ up to a normalization factor $1/\sum_i s_i$. If we considered the most regular and uniform case — a completely equidistant distribution of data, yielding all equal spacings $(1/(n + 1))$ — then we want the value of our test statistic for such a configuration to be an extermum of its support. This is achieved by including a normalization at each layer of RPS. Doing so ensures that the equidistant samples remain equidistant in each layer, thus summing over the minimal contributions to the Moran test, which then yields the smallest possible RPS value. This minimum value of RPS(n), given by the configuration of equidistant samples, can be expressed easily, as each spacing $s_{i,j}$ is equal to $\frac{1}{n+2-j}$, and thus:

$$\mathrm{RPS}_{\min}(n) = -\sum_{j=1}^{n} \sum_{i=1}^{n+2-j} \log\left(\frac{1}{n+2-j}\right) = \sum_{j=1}^{n} (j+1) \cdot \log(j+1). \tag{2.6}$$

At the other extreme, very small spacings will yield a large contribution to the sum of eq. (2.3), thus $\max(\text{RPS}(n)) = \infty$ for any given number of samples n. These extrema show that RPS measures the irregularity in sample positions. The RPS statistic increases the more samples aggregate into local clusters.

The RPS quantity calculated so far has an infinite support. We transform the RPS into a new quantity RPS*, with support $[0, 1]$:

$$\text{RPS}^*(n) = \frac{\text{RPS}_{\text{min}}(n)}{\text{RPS}(n)} \tag{2.7}$$

since the bounded interval makes extending the approximating function to the extrema of the test's support easier. This is the definition that we consider when using the RPS test and for the remainder of this note. An interesting property of the construction of RPS is that spacings in the middle (order-wise, not w.r.t. the analysis window) will have a larger impact on the overall value of the statistic compared to spacings towards the edges: this means that the test is more sensitive to centrally located non-uniformities. Such a behaviour is not uncommon, in fact both the KS and AD tests do not posses uniform sensitivity over the analysis window: KS is more sensitive towards the middle while AD is more sensitive towards the edges.

The following pseudo code (algorithm 1) illustrates how the computation of the RPS value can be implemented:

```
$x = [0, x_{(1)}, x_{(2)}, \ldots, x_{(n)}, 1]$
$rps = 0$
$min\_rps = min\_rps\_function(n)$                          ▷ see eq. (2.6)
$s = x[\texttt{first} + 1 : \texttt{last}] - x[\texttt{first} : \texttt{last} - 1]$    ▷ initial spacings
while $len(s) > 1$ do
    $rps = rps - sum(log(s))$
    $s = s[\texttt{first} : \texttt{last} - 1] + s[\texttt{first} + 1 : \texttt{last}]$   ▷ spacings for next iteration
    $s = s/sum(s)$                                          ▷ normalize
$normalized\_rps = min\_rps/rps$
```

Algorithm 1. Calculates the recursive product of spacings rps from ordered samples $x_{(i)}$.

This algorithm has a computational complexity of $O(n^2)$, and can become inefficient for very large sample sizes n. In this work we limit ourselves to $n \leq 1000$.

In an analogue way, we can also define an extension to Greenwood $G(n)$, that instead of logarithms of spacings, sums over the squares of spacings. This means that we substitute eq. (2.3) with $G_j = \sum_{i=1}^{n+2-j} \left(s_{i,j}\right)^2$, while keeping the definition of $s_{i,j}$ from eq. (2.5). We call this recursive form the "RSS" test statistic in the following comparison.

2.1 Illustration

To illustrate better how our test statistic works, and to highlight differences to other tests, we use the example set of samples drawn from a uniform (null hypothesis H_0) and a non-uniform distribution, respectively, shown in figure 2. The example given is a particularly challenging one and is used to illustrate the workings of different tests and highlight their difference, but it is not meant as a

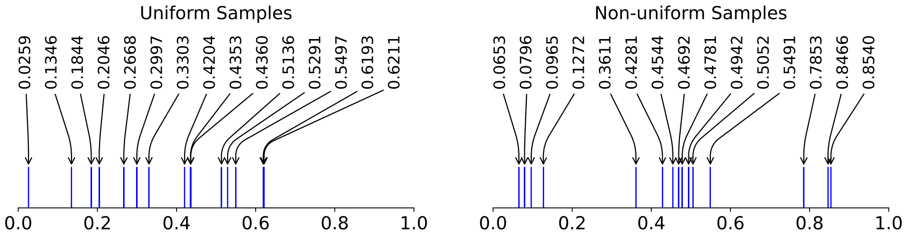


Figure 2. Example of 15 standard uniformly distributed samples (left) and 10 standard uniformly + 5 normally ($\mu = 0.5, \sigma = 0.1$) distributed samples (right). The sample positions on the $[0, 1]$ interval are annotated via the arrows + text.

performance comparison between them. Actual performance comparison using a large number of random replications are given in the following chapters.

The Moran test is based on the spacings between samples, and the smallest and largest spacings in the specific example are present in the uniform case. This leads to a more extreme test statistic value t and hence p-value $p = P(T \geq t|H_0 = \mathcal{U}(0, 1))$ of 0.117 for the uniform case, while it evaluates to $p = 0.335$ in the non-uniform case.

The KS test can detect such clustering via the CDF, however in our chosen example it is challenged by the fact that samples trend towards the left in the uniform case, while they are more balanced in the non-uniform case. This leads to p-values of 0.048 for uniform, and 0.356 for non-uniform, respectively.

The RPS test, however, taking into account also higher order spacings, finds a p-value of 0.532 for the uniform case, and a much lower p-value of 0.057 for the non-uniform samples. The behaviour of RPS is further illustrated in figure 3, that shows the individual contribution of spacings of all recursion levels that build up the test statistic value. The Moran statistic corresponds to the sum over the first row (M_1), while all subsequent levels are added for RPS. By construction, Moran's test does not preserve information about the position of spacings, meaning that the value of the test is unchanged under reordering of spacings (the test's defining is invariant due to the commutative property of sums and products): clusters of samples, as in the non-uniform case, do not affect Moran's test. Including the recursive layers allows to preserve the information about relative position of small spacings. This can be noticed by the stronger contributions to the RPS test value coming from different layers in the presence of a cluster of events (darker color on the right panel of figure 3) opposed to the small contributions coming from layers beyond the first one in the case of uniform events (left).

3 Cumulative distribution of RPS

In order to use RPS as a statistical test yielding p-values, we need its cumulative distribution F. In the case of $n = 1$ that has only two spacings — the simplest non-trivial case we can encounter — the distribution of the only events present is the standard uniform. So it is possible to write the formula of the test as a function of the sample value and find its distributions $\text{RPS}^*(1)$ as a simple transformation of random variables, which is:

$$F_{\text{RPS}^*}(x; n = 1) = 1 - \sqrt{1 - 4^{\frac{x-1}{x}}} \tag{3.1}$$

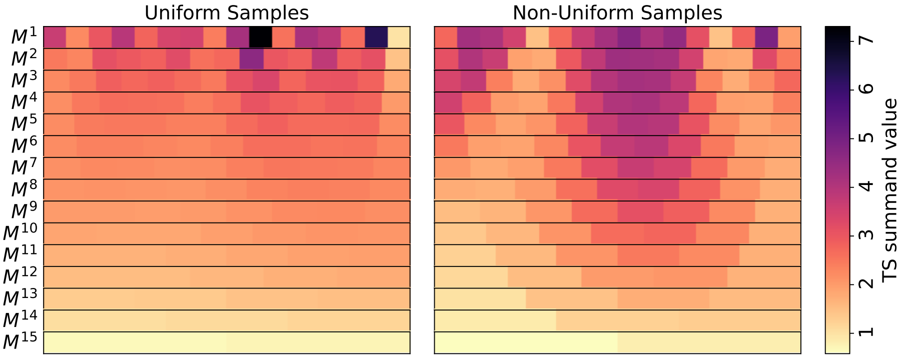


Figure 3. Illustration of the test statistic contributions from all recursion levels for the uniformly distributed samples (left) and the non-uniform samples (right). The sum over the first level only (M_1) is equivalent to the Moran statistic.

For $n \geq 2$, however, it is not simple to derive this distribution. Therefore, we resort to numerically approximating the distribution of RPS* discussed in the following section.

3.1 Approximate distribution

We have built an approximation for the cumulative distribution $F_{\text{RPS}^*}(x; n)$ precise enough to compute meaningful p-values up to relatively extreme values of up to 10^{-7}, and large sample sizes n of up to 1000. Figure 4 shows some examples of RPS* distributions for a few values of n.

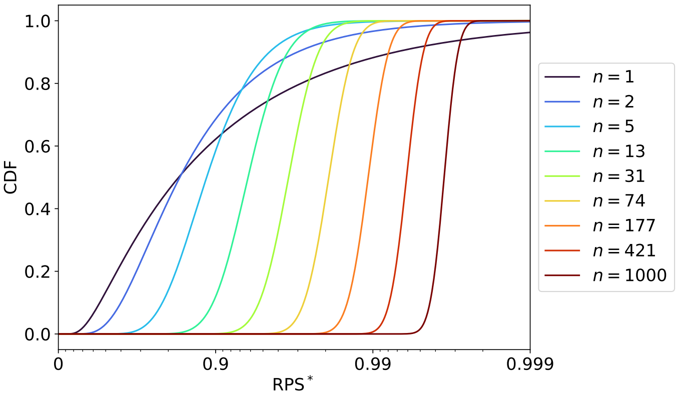


Figure 4. Example of CDFs of the RPS* distribution for a few different values of n. N.B.: the x-axis is displayed in *inverted* logarithm.

We base our approximation on simulation, drawing events with uniform distribution in the range $[0, 1]$ for a given n, and collecting $N = 2 \cdot 10^8$ samples of RPS$^*(n)$. Such simulation could be directly used to calculate p-value estimates by counting the fraction of trials below or above an observed RPS* value x for a fixed n. However, we want to provide a continuous and smooth function valid for any $n \leq 1000$. For this, we use simulated data to infer the values x of our test statistics corresponding to a discrete list of specific quantiles $p \in [10^{-7}, 1 - 10^{-7}]$. Taking the i-th element in the sorted simulation set gives an estimate for the value of $x(p = i/N)$. In order to improve this estimate, we could use bootstrapping [18], collecting different realisations of x by resampling the original dataset with replacement, resulting in a distribution of values of x for each p, from which we can then extract the mean and the standard deviation, indicative of the error (see figure 5). Instead of manually performing the bootstrapping, we can calculate the probability of each sample x to represent a specific quantile p if we were to sample randomly with replacement. For simplicity, let us consider rational quantiles that can be expressed in the form $p = \frac{k}{N}$; the probability that the i-th sample could end up representing the k-th quantile is:

$$\pi_{k,i} = F_{\mathrm{B}}\left(k, N+1-k; \frac{i}{N}\right) - F_{\mathrm{B}}\left(k, N+1-k; \frac{i-1}{N}\right) \tag{3.2}$$

where $F_{\mathrm{B}}(a, b; t)$ is the cumulative function of the Beta distribution with parameters $[a,\ b]$ estimated at t. The distribution $\mathrm{Beta}(k, N+1-k)$ represents the k-th order statistic of the uniform distribution [19], i.e. the k-th largest element of a set on N uniformly distributed random variable. Eq. (3.2) corresponds to the limiting case of performing an infinite number of bootstrapping steps and can be used to quickly estimate the mean and standard deviation of all $x(p)$ for a choice on n, especially when dealing with large datasets:

$$\mathrm{E}\left[x\left(\frac{k}{N}\right)\right] = \sum_{i=1}^{N} x_i \cdot \pi_{k,i} \tag{3.3}$$

$$\mathrm{Std}\left[x\left(\frac{k}{N}\right)\right] = \sqrt{\sum_{i=1}^{N}\left(x_i - \mathrm{E}\left[x\left(\frac{k}{N}\right)\right]\right)^2 \cdot \pi_{k,i}} \tag{3.4}$$

It would be inefficient to produce such simulation for any n, and hence we repeat the above procedure for only 180 different choices of n between 2 and 1000 following approximately a logarithmic spacing.

3.2 Fitting procedure

Using eq. (3.3) and eq. (3.4) we are able to define a grid of points with mean $\mu(n, p)$ and standard deviation $\sigma(n, p)$. Our goal is to estimate a set of points $\hat{x}(n, p)$, which will be the basis to interpolate and infer the distribution of the test statistic for all values of n and p defined above. The points $\hat{x}(n, p)$ is allowed to deviate from the means $\mu(n, p)$ within the uncertainties $\sigma(n, p)$, and can thereby provide a more accurate approximation by smoothing out stochastic noise. Additionally, points from the analytic solution for $n = 1$ (eq. (3.1)) are added to the list as anchor points at the boundary.

Given a trial set $\tilde{x}(n, p)$, we interpolate a cubic spline polynomial across the values of n for each value of p, similarly to the fits shown in figure 5. Given one such cubic spline, we evaluate the third derivative on both sides of each node, calculating the square of their difference and summing up across all nodes. Since we are using cubic splines, the third derivative is not continuous, and the "size" of the discontinuity is indicative of the smoothness of the interpolation. Summing up the contributions form all nodes of all cubic splines construct the smoothing cost function. The construction of this cost function is based on [20–22], where smoothness is treated very similarly. The estimation of the cubic spline coefficients and the evaluation of the smoothness cost function can be represented as a quadratic objective function, which we want to minimize:

$$G(\tilde{x}) \propto \frac{1}{2}\tilde{x}^T \cdot Q \cdot \tilde{x} + \bar{h}^T \cdot \tilde{x} \tag{3.5}$$

In addition to obtaining a smooth fit, there are also some additional constraints that need to be considered: monotonicity and sum of squared residuals.

Since the samples $\tilde{x}(p|n)$ should represent a cumulative density function, then it is important they are properly ordered, ensuring that $\tilde{x}(p_i|n) \leq \tilde{x}(p_j|n)$ for $i \leq j$. This is ensured including a number of linear inequality constraints modelled as a linear constraint matrix:

$$A \cdot \tilde{x} \leq b \tag{3.6}$$

Lastly, we assume that the values $\tilde{x}(n,p)$ are normally distributed with means $\mu(n,p)$ and standard deviations $\sigma(n,p)$. Since we want to move away from the initial values $\mu(n,p)$ in order to obtain a smoother fit, it is important to limit this movement the further away we get and we do so by considering the sum of squared residuals, which is a typical measure to account for the global deviation from the mean. Since we assume gaussian deviations, the sum of all squared residuals can be modelled by a χ^2 distribution with m degrees of freedoms, where m is the total number of parameters, i.e. the number of nodes. Given this distribution, we can estimate the value of the cost function to be limited to the mean (m) plus one standard deviation ($\sqrt{2m}$) of the χ^2 distribution, thus:

$$\sum_{i=1}^{m} \frac{(\tilde{x}_i - \mu_i)^2}{\sigma_i^2} \leq m + \sqrt{2 \cdot m} \tag{3.7}$$

Figure 5 shows a fitted spline representation of $\hat{x}(n|p)$ for different values of p. Based on the resulting list of corresponding p and $\hat{x}$ values, that we obtained for any n, we generate another spline interpolation as the approximation of the desired cumulative distribution $F(\hat{x}; n)$ for a given n. As the cumulative distribution function F is strictly monotonous in $\hat{x}$, we use the [23] monotonic spline interpolation on the points $[\hat{x}(p|n), p]$ to produce the final CDFs, shown in figure 4 for a few values of n.

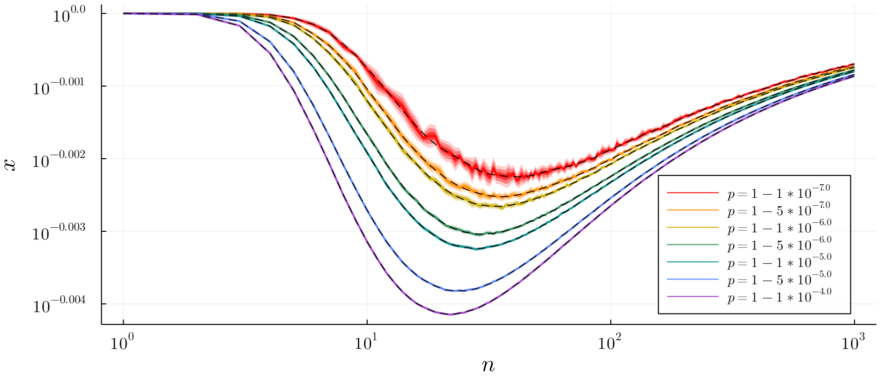
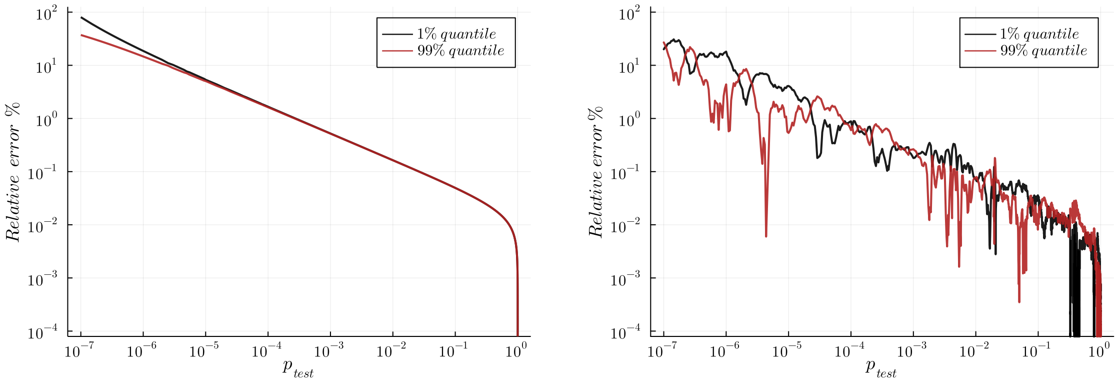


Figure 5. Example of spline fitted x-values across n for a few extreme p-values. The colored bands show the 1, 2 and 3 sigma bands estimated via bootstrapping, the black, dashed lines show the approximations by the spline fits.

3.3 Error estimation

Finally, we are also able to estimate the precision of our approximation. Given any set of i.i.d. random variables, such as x, the corresponding list of estimated quantiles p represents a random set of uniform variates. For any rational quantile $p_{\text{test}} = \frac{k}{N}$ we can estimate the 98% credible interval $(p_{0.01}, p_{0.99})$ using the distribution of the k-th order statistic $\text{Beta}(k, N + 1 - k)$. Given the credible interval, we calculate the relative error of p_{test} against the extrema of the interval, considering the largest value representative of the relative error of a random ECDF up to a specified credible level. The results of the estimated relative error for our choice of $N = 2 \cdot 10^8$ and for quantiles as low as $p = 10^{-7}$ are shown in figure 6(a).



(a) Estimated relative error of empirical p-value with respect to the 98% credible interval and $2 \cdot 10^8$ samples. The vertical axis reports the scale of the relative error in percent for two extremes, the 1% and the 99% quantile of the order statistic distribution.

(b) Estimated relative error of fitted p-value with respect to p-values obtained via bootstrapping. The vertical axis reports the scale of the relative error in percent for two extremes, the 1% and the 99% quantile of the bootstrapping distribution. Results for $n = 75$.

Figure 6. Per-quantile relative error stimation of the approximate RPS distribution.

As expected, the errors are increasing towards smaller p-values and exhibit an approximately linear behaviour in the log-log plot. We see that the estimated upper bound of the relative error for a p-value of 10^{-3} is below 1%, while for a p-value of 10^{-5} it increases to $< 10\%$ and ultimately to $< 100\%$ for p-values of 10^{-7}. Such a "large" relative error for small p-values may sound alarming at first, but estimating a p-value of 10^{-7} and knowing it could actually be closer to $2 \cdot 10^{-7}$ would hardly change the statistical interpretation of a result.

In order to show the validity of these results, we compute the relative error of our approximate distributions against a test dataset containing 10 times more samples using bootstrapping. We do so for a few choices of number of events n, and in figure 6(b) it can bee seen that the behavior of the relative error is in complete agreement with our analytic estimates of figure 6(a).

So defined, the relative error $\delta(p|N)$ is a function of the quantile p and number of samples N, but this relationship can also be inverted in order to determine the number of samples necessary to achieve a desired relative error for a specific quantile: $N(p|\delta)$. Our choice of $N = 2 \cdot 10^8$ was in fact guided by the requirement of having a relative error lower than 100% for a p-value of 10^{-7} in at least 99% of cases.

It is worth stressing that these estimates of the relative error are accurate with respect to the ECDF that was sampled for each independent n, but might be subject to small changes after the smoothing fit we performed in order to regularize and infer the distributions for all missing values of n.

3.4 Implementation

The RPS test is made available as open-source packages for Python[1] and Julia,[2] respectively, with the p-value parametrizations initially available up to 1000 samples.

Below we give a minimal example to evaluate the RPS test for an array x in both language implementations, with x being:

```
x = [0.1, 0.4, 0.76]
```

The python library can be used like the following:

```
>>> from spacings import rps
>>> rps(x, "uniform")
RPStestResult(statistic=0.9547378863245608, pvalue=0.8865399970192409)
```

and the Julia equivalent giving identical results in the following:

```
>>> using SpacingStatistics, Distributions
>>> rps(x, Uniform())
(statistic=0.9547378863245608, pvalue=0.8865399970192409)
```

4 Example application 1: bump hunting

In this section, we illustrate how the RPS test could be used in a physics scenario. We consider a detector that collects a number of events in an observable x, where x could for example be the energy of an event, the detection time, or a reconstructed quantity like an invariant mass. We expect some or all of the observed events to follow a known background distribution $f_{\mathrm{B}}(x)$, but there may be an additional contribution of events from an unknown signal distribution $f_{\mathrm{S}}(x)$— such as a rare, exotic particle decay with unknown mass. Hence we want to quantify the goodness-of-fit of the background only model to our data. A resulting low p-value could indicate the presence of events distributed according to an additional, unknown signal distribution.

In the example here, we use an exponential distribution $f_{\mathrm{B}}(x) = e^{-x}$ for the background model (null-hypothesis). In order to illustrate how the presence of an actual signal (alternative hypothesis) would affect the outcome, we also inject additional events following a normal distribution centred at $x = 1$ and width $\sigma = 0.05$. The number of events is Poisson fluctuated for both background and signal, with expected values of $\langle n_{\mathrm{b}} \rangle = 100$ and $\langle n_{\mathrm{s}} \rangle$ varied as specified. In figure 7, an example distribution of observed events is shown, together with the assumed background distribution, and the distribution with injected signal (here $\langle n_{\mathrm{s}} \rangle = 5$).

The example case chosen is similar to that, for instance, of a search for an exotic particle with unknown mass — a problem sometime referred to as “bump hunting”. In this case, x would represent an invariant mass.

[1] https://pypi.org/project/spacings/.

[2] https://github.com/bat/SpacingStatistics.jl/tree/dev.

N.B., we do not assume that we know the rate of the underlying processes, meaning that the number of observed counts is not included in our analysis other than for the calculation of the test statistic. This means that we test for the "shape" of the distribution, not its normalization. The conversion of events via the CDF of the distribution under test f_B transforms the problem into a test of uniformity.

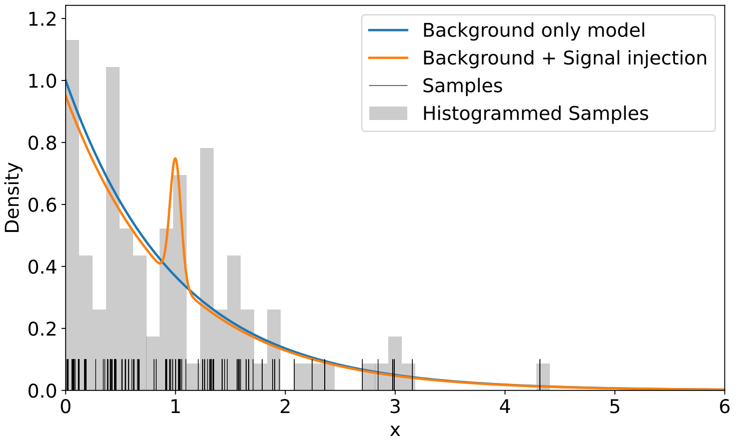


Figure 7. Example physics problem, with observed events distributed in x. We test the goodness-of-fit of the background only model (blue) to the samples. Here the samples have been generated according to a different distribution with an injected signal (orange).

The p-value distributions under the assumption of H_0 (i.e. only background is present) for repeated trials with $\langle n_\mathrm{b}\rangle = 100$, and various injected $\langle n_\mathrm{s}\rangle = [0, 3, 6, 9, 12, 15]$ are shown in figure 8. All distributions with no signal ($\langle n_\mathrm{s}\rangle = 0$) show a flat p-value distribution as expected, since in that case all events are drawn from the background distribution p_B. For trials with injected signal, the distributions are trending towards smaller p-values, indicating the worsened goodness-of-fit for the background only model. In the example, all tests exhibit this behaviour, while the RPS test offers the largest rejection probability of the null hypothesis.

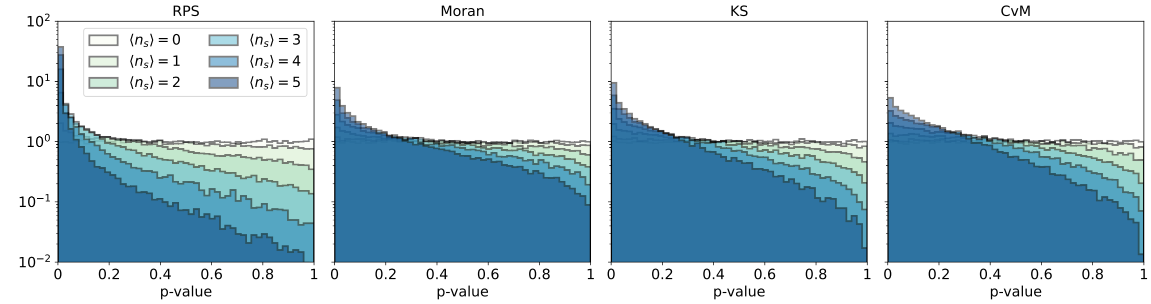


Figure 8. P-value distributions for background only samples ($\langle n_\mathrm{s}\rangle = 0$) and background plus randomised signal injections comparing to the background model for several choices of test statistics.

We quantify the sensitivity of the analysis to reject the background only model at different significance levels under the assumption of the presence of a signal. Therefore we check the median p-value of repeated trials, and at what value of $\langle n_\mathrm{s}\rangle$ it crosses specific critical values (see left panel

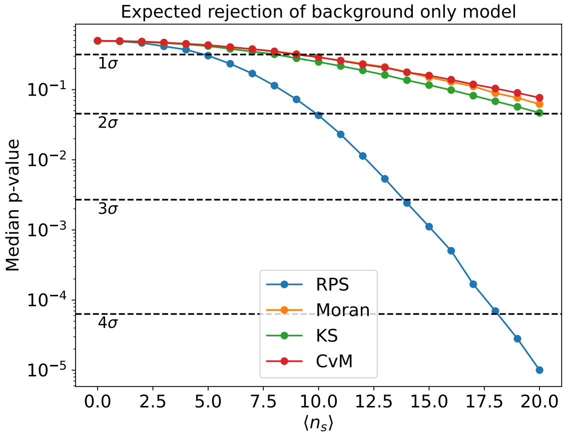


Figure 9. The expected significance level at which the background model can be excluded under the assumption of a signal, as a function of $\langle n_\mathrm{s} \rangle$ for the different tests.

of figure 9). In our chosen example, for a signal of strength $\langle n_\mathrm{s} \rangle = 10$ we expect to reject the background only model using RPS at the 2σ significance level,[3] whereas for the other tests, a signal of at least $\langle n_\mathrm{s} \rangle = 20$ is needed to achieve the same. Such a large signal of $\langle n_\mathrm{s} \rangle = 20$ would allow to reject the background only model at $> 4\sigma$ significance with the RPS test.

5 Example application 2: trigger for transient neutrino emission

This section summarizes one of the first applications of the RPS test in astrophysics, namely for triggering transient events in cryogenic neutrino detectors [24] such as the RES-NOVA experiment [25]. While the technical details about the experimental setup, the simulation and the application of the RPS test can be found in the aforementioned references, here we will summarize some highlights.

Cryogenic neutrino detectors can be described as counting experiments, that output a temporal data stream of observed neutrino interactions. Without the presence of a transient neutrino source, we only expect some activity from background events. If a source of neutrinos is placed at an observable distance, such as a core-collapse supernova (CC-SN) at 10 kpc, we expect a short burst of neutrinos resulting in an excess in the observed counts over the background only expectation. The sources of such transient neutrinos can vary in their overall duration, temporal distribution and amplitude. Figure 10 shows as examples the expected counts of two different neutrino sources, a CC-SN and a failed CC-SN, respectively, together with a constant background expectation at a rate of 0.18 Hz.

To issue alerts in near real time about the presence of such sources, one needs a triggering system with a chosen false alarm rate (FAR), which is set to 1 per week for SNEWS [26]. The standard approach for building such triggers is the usage of Poisson statistics, that analyse the data stream in windows of a fixed length, and check the level of observed counts compared to the expectation from background, see for example ref. [27]. The Poisson approach works well if the window size is chose optimally for a given signal. However, if the chosen window size does not match the signal,

[3] A significance level in terms of numbers of k standard deviations σ can be translated to a p-value as one minus the integral over a unit normal distribution form $-k$ to $+k$.

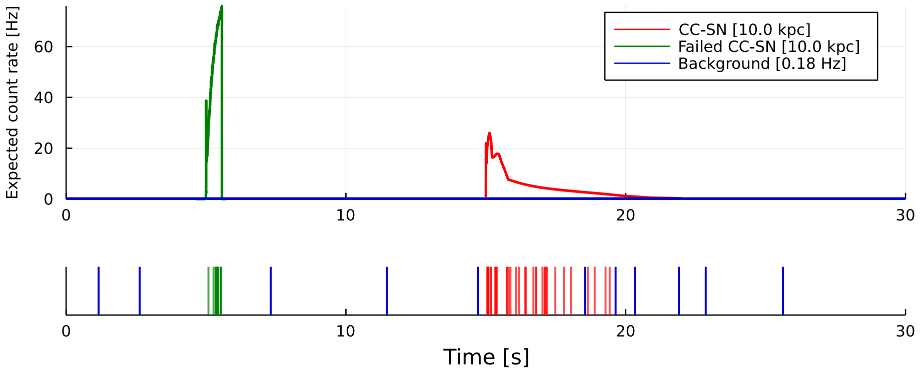


Figure 10. Example of observed counts at a neutrino detector for signals from a core-collapse SN (at time $t = 15$ s) and a failed core-collapse SN (at $t = 5$ s) for progenitors stars with 27 $M_\odot$ and 40 $M_\odot$ respectively, both at a distance of 10 kpc. (Modified version of a figure from ref. [24].)

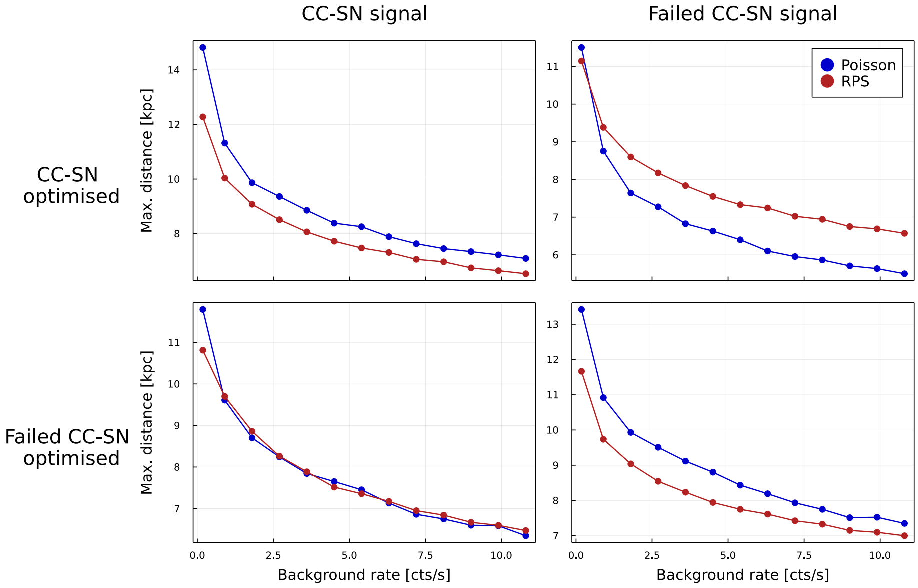


Figure 11. Maximum distance probed at a 95% success rate as a function of time with respect to two sample signals, the Core-Collapse SN and the failed Core-Collapse SN, obtained using analysis windows optimised on each of the tested signals. (Figure from ref. [24].)

the performance is affected as either the signal is not contained in the window (window too small), or the window is too large and the signal is washed out by background events. Performance of the Poisson test, as a function of the background rate and for the case of optimal window choice for the CC-SN and the failed CC-SN signals is shown in figure 11.

The RPS test can likewise be used to analyze data streams to look for transient phenomena. Here we do not make any explicit assumption on the background rate, but rather assume that the background is constant in rate, which means the distribution of background events in the time dimension is following a uniform distribution. With RPS we can test for this uniformity, which can be used to detect short additional contribution of events in the data. The performance of the RPS test as trigger is also shown in figure 11.

In the case where the window size for the Poisson test is optimal, the performance can not be matched with RPS (the panels in the upper left and lower right, respectively, in figure 11) and results in up to 10% lower sensitivity. However, the more interesting case is when using the window size optimized for one signal for the analysis of a different signal (the panels in the upper right and lower left, respectively, in figure 11). The RPS test is more robust to such changes, and in the example of searching for a failed CC-SN signal with a window optimized on a particular CC-SN scenario, we find up to 20% increase in sensitivity.

In general, what we find is that the RPS test being non parametric and able to deal with much larger analysis windows is more robust to changing conditions. Less assumptions about the background rate and the expect signals have to be made at the trade off of being non-optimal to one specific signal choice, but good performance for the more agnostic case of unknown signal distributions. This makes RPS an interesting choice for a general-purpose, agnostic trigger algorithm for the search of transient events.

6 Performance comparison

This section presents an in-depth performance comparison of the RPS test to several other tests referenced in the introduction (KS, AD, CvM and Moran — all those that allow to compute p-values). We are interested in detecting small changes in an otherwise uniform distribution, and therefore construct the following generic benchmark scenario: for one simulation of a specific test case $H^K(n, s, w)$ we generate $(1 - s) \cdot n$ random variates[4] from a standard uniform distribution $\mathcal{U}(0, 1)$, where s is a *signal* fraction. In addition, we include $s \cdot n$ samples distributed according to $\Delta + \mathcal{U}(0, w)$ with the offset $\Delta = \mathcal{U}(0, 1 - w)$, i.e. a more narrow uniform distribution of width w over a random interval within $(0, 1)$. In our comparison, we vary all three parameters of $H^K(n, s, w)$, i.e. the number of samples n, as well as the fraction s and width w of the injected *signal* events. A sensitive test should be able to detect the presence of the added, narrower signal samples by reporting a low p-value.

Figure 12 show the performance of our choice of tests as a function of the above three parameters. As a metric, we show the median p-value obtained from repeated trials, and we interpret a lower reported median p-value as a more powerful test. This number can be interpreted as the median significance at which we expect to be able to reject the null hypothesis. What can be observed is, that for all the tested scenarios the RPS test is performing either on par or significantly better than the Moran test. The ECDF based tests (KS, AD or CvM) start to dominate in terms of performance only for relatively wide signals of around 25% total width or more. When analysing the goodness-of-fit given a large number of samples, i.e. order of several hundreds, the differences between RPS and the ECDF-style tests start to become smaller. Overall, the outcome of this performance study suggests that when signals are expected of widths that span over less than a 25% percentile of the null hypothesis distribution, and if the number of samples is $n < 1000$, the RPS tests compares very favourably against all others considered.

We also investigated other metrics to judge the test's performance, such as the area under the receiver operating characteristics (ROC) curve between signal and null hypothesis trials. The overall picture does not change substantially.

[4] Numbers of samples are rounded to the closest integer.

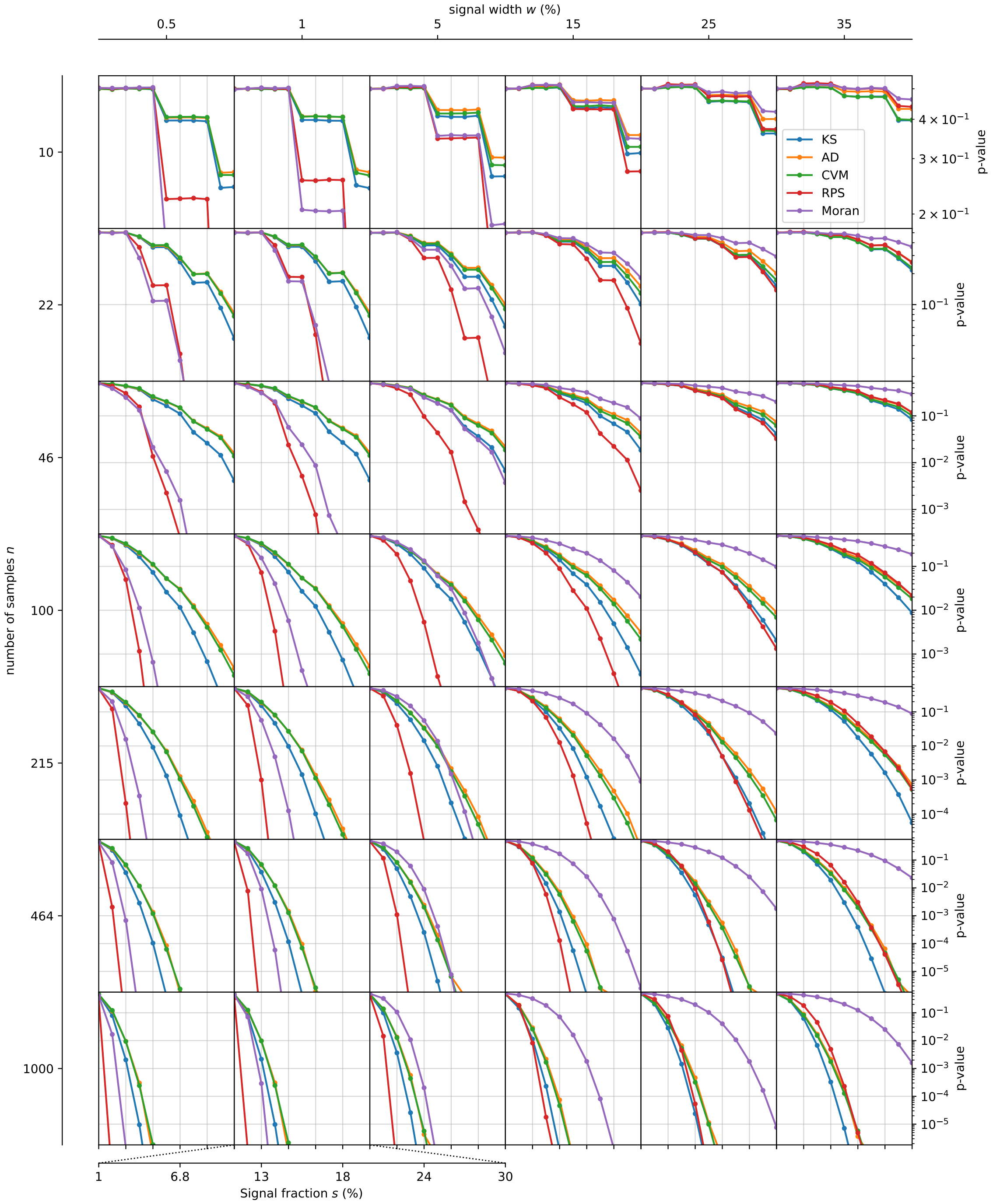


Figure 12. Comparison of the performance (median p-value of repeated trials, individual panel's xy-axis) as a function of the number of total samples (large y-axis), the width of the signal (large x-axis) samples, and the fraction of signal samples (individual panel's x-axis). The number of signal samples is rounded to the closest integer, hence the "step"-like features visible mostly in the first few rows.

7 Conclusions

The RPS test statistic is a sensitive measure to detect deviations of samples from a continuous distribution with known CDF. The analytic distribution of the RPS statistic is not available for $n > 1$, but a high accuracy parameterization valid up to sample sizes of $n = 1000$ is provided in order to use RPS as a goodness-of-fit test. In the presented test scenarios, the RPS test outperforms other tests significantly under certain circumstances, in particular when the observed sample is small ($n < 1000$) and introduced deviations are narrow, i.e. concentrated over a small quantile. Two example physics analysis cases were presented, we show that the sensitivity of a "bump hunting" experiment could be boosted by up to a factor of two by choosing the RPS test over others. And we show how RPS can be used to build a robust and agnostic trigger algorithm for a SN experiment.

Acknowledgments

We would like to thank Allen Caldwell, Oliver Schulz and Johannes Buchner for helpful discussions and comments.

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy — EXC-2094 — 390783311.

References

[1] A. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione*, *G. Ist. Ital.* **4** (1933) 83.

[2] N. Smirnov, *Table for Estimating the Goodness of Fit of Empirical Distributions*, *Ann. Math. Stat.* **19** (1948) 279.

[3] T.W. Anderson and D.A. Darling, *A Test of Goodness of Fit*, *J. Am. Statist. Assoc.* **49** (1954) 765.

[4] Y. Marhuenda, D. Morales and M.C. Pardo, *A comparison of uniformity tests*, *Statistics* **39** (2005) 315.

[5] K. Pearson, *Note on Francis Galton's problem*, *Biometrika* **1** (1933) 390.

[6] K. Pearson, *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** (1933) 379.

[7] H. Cramér, *On the composition of elementary errors*, *Scand. Actuarial J.* **1928** (1928) 13.

[8] R. von Mises, *Wahrscheinlichkeit Statistik und Wahrheit*, Springer-Verlag (1928).

[9] R. Pyke, *The Supremum and Infimum of the Poisson Process*, *Ann. Math. Stat.* **30** (1959) 568.

[10] J. Durbin, *Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals*, *Biometrika* **56** (1969) 1.

[11] H.D. Brunk, *On the Range of the Difference between Hypothetical Distribution Function and Pyke's Modified Empirical Distribution Function*, *Ann. Math. Stat.* **33** (1962) 525.

[12] R.C.H. Cheng and M.A. Stephens, *A goodness-of-fit test using Moran's statistic with estimated parameters*, *Biometrika* **76** (1989) 385.

[13] M. Greenwood, *The Statistical Study of Infectious Diseases*, *J. Roy. Stat. Soc.* **109** (1946) 85.

[14] L. LeCam, *Un théorème sur la division d'un intervalle par des points pris au hasard*, *Publ. Inst. Statist. Univ. Paris* **VII** (1958) 7.

[15] N. Cressie, *On the Logarithms of High-Order Spacings*, *Biometrika* **63** (1976) 343.

[16] N. Cressie, *An Optimal Statistic Based on Higher Order Gaps*, *Biometrika* **66** (1979) 619.

[17] L. Shtembari and A. Caldwell, *On the sum of ordered spacings*, arXiv:2008.02048.

[18] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, *Annals Statist.* **7** (1979) 1.

[19] H.A. David and H.N. Nagaraja, *Order statistics*, Wiley (2003).

[20] P. Dierckx, *An algorithm for smoothing, differentiation and integration of experimental data using spline functions*, *J. Comput. Appl. Math.* **1** (1975) 165.

[21] P. Dierckx, *A Fast Algorithm for Smoothing Data on a Rectangular Grid while Using Spline Functions*, *SIAM J. Numer. Anal.* **19** (1982) 1286.

[22] P. Dierckx, *Curve and surface fitting with splines*, in *Monographs on numerical analysis*, Clarendon Press (1996).

[23] F.N. Fritsch and J. Butland, *A Method for Constructing Local Monotone Piecewise Cubic Interpolants*, *SIAM J. Sci. Stat. Comput.* **5** (1984) 300.

[24] P. Eller, N. Iachellini Ferreiro, L. Pattavina and L. Shtembari, *Online triggers for supernova and pre-supernova neutrino detection with cryogenic detectors*, *JCAP* **10** (2022) 024 [arXiv:2205.03350].

[25] RES-NOVA collaboration, *RES-NOVA sensitivity to core-collapse and failed core-collapse supernova neutrinos*, *JCAP* **10** (2021) 064 [arXiv:2103.08672].

[26] SNEWS collaboration, *SNEWS 2.0: a next-generation supernova early warning system for multi-messenger astronomy*, *New J. Phys.* **23** (2021) 031201 [arXiv:2011.00035].

[27] N.Y. Agafonova et al., *On-line recognition of supernova neutrino bursts in the LVD detector*, *Astropart. Phys.* **28** (2008) 516 [arXiv:0710.0259].