Is the bump significant? An axion-search example

Frederik Beaujean^{1,a} , Allen Caldwell^{2,b} , Olaf Reimann^{2,c}

¹ C2PAP, Excellence Cluster Universe, Ludwig-Maximilian University of Munich, Munich, Germany

² Max Planck Institute for Physics, Munich, Germany

Received: 21 March 2018 / Accepted: 5 September 2018 / Published online: 28 September 2018
© The Author(s) 2018

Abstract Many experiments in physics involve searching for a localized excess over background expectations in an observed spectrum. If the background is known and there is Gaussian noise, the amount of excess of successive observations can be quantified by the SQUARES (squares-of-run-residuals sum) statistic taking care of the look-elsewhere effect. The distribution of the SQUARES statistic under the background model is known analytically but the computation becomes too expensive for more than about a hundred observations. Here a principled high-precision extrapolation from a few dozen up to millions of data points is derived. It is most precise in the interesting regime when an excess is present. The method is verified for benchmark cases and successfully applied to real data from an axion search. C++ code that implements our method is available at <https://github.com/fredRos/runs>.

1 Introduction

Searching for a bump at an unknown location in a spectrum is a common task faced in analysis of data. In this paper, we are interested in the case where the data are assumed to fluctuate according to a Gaussian distribution and the background expectation is known. Minimal assumptions are made on the signal shape - basically only that it is seen as a contiguous excess above the background expectation. We have previously defined a test statistic and derived an analytic expression for its probability distribution [1], fully taking into account the trials factor (look-elsewhere effect). We extend our previous work by deriving approximate distributions for the test statistic applicable to large data sets.

Specifically we assume there are L observations $\{y_i\}$ and the index i represents an ordering of the data such as time,

mass, energy.... We assume a background model that has no signal. In this model, the observations independently follow a Gaussian or Normal distribution

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i), \quad (1)$$

where the expectation μ and the standard deviation σ are known for every i . In our previous work [1], we introduced a test statistic T to check the consistency of the background model with the observations. If a discrepancy is found, more specific analyses can be carried out to decide if a signal is present and to determine the parameters describing the bump in a separate signal model.

We focus on the case when no explicit signal model is available. The only property of such a model relevant to the present discussion is that, if present, a signal produces a bump or peak; i.e., a localized excess over the background model. The most common test statistic used in this setting is the classic χ^2 statistic:

$$\chi^2 = \sum_{i=1}^L \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2. \quad (2)$$

Although χ^2 is considered a good omnibus test statistic (i.e., good power against a variety of alternative models), it has an obvious limitation for bump hunting. The value of χ^2 is independent of the order of observations; i.e., any permutation of the observations yields the same value of χ^2 . Yet in informal terms, a bump is a clustering of large positive residuals that come one after another, so the ordering is crucial.

In the statistics community, statistics involving *runs*, or sequences of observations sharing a common attribute commonly called a *success*, are studied in great detail. In this work, an observation is a success if it is above the background expectation; i.e., $y_i > \mu_i$. The case when all successes are indistinguishable, or equally weighted, such as in Bernoulli trials has received a lot of attention; we refer to [1] for a brief survey of the relevant literature. Common statistics based on

^a e-mail: Beaujean@mpp.mpg.de

^b e-mail: Caldwell@mpp.mpg.de

^c e-mail: Reimann@mpp.mpg.de

runs are the number of runs, or the length of the longest run. It is known that these statistics are, at least asymptotically, independent of the χ^2 statistic; i.e., they provide “orthogonal” information; see [2, Ch. 11]

A simple approach to combine χ^2 and runs is to compute a p value for each statistic and somehow combine the two into a single p value. However, this involves many subtle choices so the result is rather arbitrary [3].

The motivation for the statistic, T , introduced in [1] is to combine the information in both the size and the order of residuals into a single test statistic, which we will refer to in this article as the SQUARES (squares-of-run-residuals sum) statistic (see Sect. 2). The SQUARES statistic is more powerful in highlighting discrepancies between data and the background model in case there is a signal peak. The superior power of SQUARES compared to χ^2 can be verified for an explicit signal model as done in [1] but for the most interesting case of an unspecified signal, the power of T is undefined.

In this article, we assume the background model known which allows us to derive expressions analytically. This ideal case is hardly ever met in practice where nuisance parameters in the background model have to be inferred from the same data for which the test statistic is computed. This leads to a biasing of the p value towards higher values. The effect has been studied quantitatively for a prototypical peak-fitting example from physics comparing T with χ^2 among other statistics in [4]. The bias in χ^2 can be removed exactly under certain conditions but these are hard to ensure in practice [4]. The bias in the p value for the SQUARES statistic becomes negligible when the number of fitted parameters is much smaller than the number of observations L . In general, since we focus here on large L , the bias in either statistic is usually not important.

A more important aspect than the bias is the look-elsewhere effect (also called the trials factor) that arises in many methods, for example in the search for the Higgs boson at the LHC [5,6]. There, the profile-likelihood ratio statistic was employed [7] which requires fully specifying both the background and the signal model including dependence on unknown parameters to be estimated from the data. For reliable estimates of the look-elsewhere effect, asymptotic normality and principled extrapolation from small to large significance had to be used [8,9]. In contrast, the probability distribution for the SQUARES statistic includes this effect and does not rely on asymptotic normality.

In this article, we derive new expressions in order to approximate the distribution of the SQUARES statistic when L is large. This is exactly the regime in which the bias is small and where the exact distribution is too expensive to evaluate due to a nearly exponential growth of compute effort as a function of L . The motivation is an axion-search experiment [10], where we eventually expect to have of order

$L = 5 \times 10^7$ power measurements. The axion signal is expected to be very narrow but with unknown shape, and we wish to make the minimum number of assumptions in a first pass through the data. We intend to use the SQUARES statistic to identify candidate signals in this spectrum. Once a candidate signal is identified, the experimental setup can be modified to increase the signal-to-noise ratio considerably. However, changing the setup and acquiring more data is time consuming, and can only be performed relatively rarely. It is therefore important to understand the statistical significance of a putative signal.

After defining the SQUARES statistic in Sect. 2, we derive the approximate distribution of the SQUARES statistic in Sect. 3 and present our investigations on numerical properties in Sect. 4. The application of SQUARES in searching for an axion signal is described in Sect. 5, and finally we conclude in Sect. 6.

2 The SQUARES statistic

In this section, we apply the general definition of the SQUARES statistic from [1] to observations with Gaussian uncertainties and we illustrate the definition with a concrete example.

Consider the sequence of L observations as consisting of success and failure runs, where the observation i is a success if it is above the background expectation, $y_i \geq \mu_i$, and a failure if $y_i < \mu_i$. A run consists of one or more successes and we assume there is at least one success. For bump hunting, we focus on successes but by reversing the role of successes and failures, one can search for dips in spectra.

To each success, we associate a weight given by the square of the residual. Then we sum the weights within a success run to assign a weight to the run, which is just the ordinary χ^2 restricted to a run. Finally, the SQUARES statistic T is the largest weight of any success run. Let S_i denote the partial sum of squared residuals in a run. If observation i is a failure, then reset $S_i = 0$. If a_j is the index of the first and b_j is the index of the last success in run j , then the weight of success run j is

$$\chi_j^2 \equiv S_{b_j} = \sum_{i=a_j}^{b_j} \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2 \quad (3)$$

and

$$T \equiv \max_j \chi_j^2. \quad (4)$$

Using the cumulative distribution function $F(T|L)$, the p value is the tail-area probability to find T larger than T_{obs} , the value for the observed data set,

Fig. 1 A sequence of 1000 Gaussian distributed random numbers (shifted and scaled) is shown in the top plot, while the value S_i of the current success run is shown in the second plot. A zoom in the region around the largest S_i is shown below

$$p \equiv 1 - F(T_{\text{obs}}|L). \quad (5)$$

As an illustration, a sequence of $L = 1000$ independent standard Gaussian distributed random numbers (shifted and scaled) is shown in Fig. 1, while the partial sum S_i is shown in the second plot. In the case shown, the largest observed χ^2 of any run leads to $T_{\text{obs}} = 15.8$ for $a = 110$ and $b = 114$. Using our results presented below, this T_{obs} is equivalent to a p value of 0.36 which suggests good agreement with the background-only hypothesis.

The exact probability distribution for the SQUARES statistic, T , has been derived in [1], and code is available on github [11] to calculate the cumulative distribution function for any value of the test statistic in *mathematica* and *C++*. The calculation time grows rapidly with L (roughly as $\exp(\sqrt{L})/L$), and for $L \approx 200$ becomes too long for practical use on today's CPUs, even with multiple cores. We derive here a formula that allows us to use the results for moderate $L \approx 100$ to extrapolate to very large $L \gtrsim 10^6$ with high accuracy in the region of interest where T_{obs} is large such that the p value is very small. We have implemented the extrapolation formula in our code [11]. This allows the use of the SQUARES statistic in very long sequences of measurements with the correct statistical distribution for the test statistic without relying on expensive and somewhat inaccurate Monte Carlo simulations.

3 Cumulative distribution function of the SQUARES statistic for large L

We assume there is a sequence of L observations. Consider a partition $L = N_l + N_r$ into two segments, where the subscripts l and r denote the left-hand and right-hand part (see Fig. 2). Suppose that in each of the segments considered separately, the test statistics T_l and T_r are both less than T_{obs} . Then there are exactly two ways this can occur. Either $T < T_{\text{obs}}$ for the entire sequence, or there is a run that crosses the boundary (condition $\mathcal{C}$) and its χ^2 in each segment is less than T_{obs} but the combined χ^2 is the largest of any run (condition $\mathcal{M}$) and exceeds T_{obs} :

$$\begin{aligned} P(T_l < T_{\text{obs}}, T_r < T_{\text{obs}} | N_l, N_r, L) \\ &= P(T < T_{\text{obs}} | N_l, N_r, L) \\ &\quad + P(T_l < T_{\text{obs}}, T_r < T_{\text{obs}}, T \\ &\quad \geq T_{\text{obs}}, \mathcal{C}, \mathcal{M} | N_l, N_r, L) \end{aligned} \quad (6)$$

We denote by $F(T_{\text{obs}}|L) \equiv P(T < T_{\text{obs}} | N_l, N_r, L)$ the value of the cumulative distribution function of the test statistic T for a total of L observations. Since the events are assumed independent, we can factorize

$$\begin{aligned} P(T_l < T_{\text{obs}}, T_r < T_{\text{obs}} | N_l, N_r, L) \\ &= P(T_l < T_{\text{obs}} | N_l, \mathcal{N}_r, \mathcal{L}) P(T_r < T_{\text{obs}} | \mathcal{N}_l, N_r, \mathcal{L}) \end{aligned} \quad (7)$$

$$= F(T_{\text{obs}}|N_l) F(T_{\text{obs}}|N_r), \quad (8)$$

into the left and right parts and re-arrange to find

Fig. 2 Visual description of reducing a long run into two shorter runs

$$F(T_{\text{obs}}|L) = F(T_{\text{obs}}|N_l)F(T_{\text{obs}}|N_r) - P(T_l < T_{\text{obs}}, T_r < T_{\text{obs}}, T \geq T_{\text{obs}}, \mathcal{C}, \mathcal{M}|N_l, N_r, L). \quad (9)$$

For both N_l and N_r large, there typically are many runs so it is unlikely that a boundary-spanning run has the largest χ^2 . Then Eq. (9) shows that the cumulative function of the whole sequence is essentially the product of cumulative functions for each segment minus a small correction.

3.1 χ^2 distribution of runs starting at the boundary

Since the boundary is fixed and we require for the correction term that the run cross the boundary, it must have the first result on each side above the expectation. Consider the run segment on the right of the boundary with weight χ_r^2 : we can use the law of total probability to calculate the probability density as the sum of probabilities for runs of different length times the probability density for χ_r^2 given the appropriate degrees of freedom:

$$h(\chi_r^2|N_r) = \sum_{k=1}^{N_r} P(\chi_r^2|k)P(k|N_r), \quad (10)$$

where k is the length of the success run and we require at least one success. We have

$$P(k|N_r) = \begin{cases} 2^{-(k+1)}, & k < N_r \\ 2^{-N_r}, & k = N_r. \end{cases} \quad (11)$$

The reason for the +1 for $k < N_r$ is the requirement that the result is below expectation for the $k + 1^{\text{st}}$ sample, which also has probability 1/2. $P(\chi_r^2|k)$ is the usual χ^2 probability density for k degrees of freedom so that we find

$$h(\chi_r^2|N_r) = \left(\sum_{k=1}^{N_r-1} \frac{1}{2^{k+1}} \cdot \frac{(\chi_r^2)^{k/2-1} e^{-\chi_r^2/2}}{2^{k/2} \Gamma(k/2)} \right) + \frac{1}{2^{N_r}} \cdot \frac{(\chi_r^2)^{N_r/2-1} e^{-\chi_r^2/2}}{2^{N_r/2} \Gamma(N_r/2)}. \quad (12)$$

For a χ^2 from a particular run to be our test statistic T , the run must span the boundary (condition $\mathcal{C}$) and its χ^2 must be the maximum value for any run in the full L range (condition $\mathcal{M}$), so $\chi_l^2 + \chi_r^2 = T$. We calculate the probability that a

contiguous run spanning the boundary satisfies the specified conditions as

$$P(\chi_l^2 < T_{\text{obs}}, \chi_r^2 < T_{\text{obs}}, \chi_l^2 + \chi_r^2 \geq T_{\text{obs}}, \mathcal{C}, \mathcal{M}|N_l, N_r, L) = \int_0^{T_{\text{obs}}} d\chi_l^2 h(\chi_l^2|N_l) \int_{T_{\text{obs}}-\chi_l^2}^{T_{\text{obs}}} d\chi_r^2 F(\chi_l^2 + \chi_r^2|L) h(\chi_r^2|N_r). \quad (13)$$

We implement condition $\mathcal{M}$ by weighting each possible contribution $\chi_l^2 + \chi_r^2$ with the probability that this is the largest in the full range L , $F(\chi_l^2 + \chi_r^2|L)$. This is the quantity we seek to compute so we cannot evaluate this expression directly. But

$$F(T_{\text{obs}}|L) \leq F(\chi_l^2 + \chi_r^2|L) \leq F(2T_{\text{obs}}|L) \leq 1, \quad (14)$$

so as $F(T_{\text{obs}}|L) \rightarrow 1$ we can write

$$P(\chi_l^2 < T_{\text{obs}}, \chi_r^2 < T_{\text{obs}}, \chi_l^2 + \chi_r^2 \geq T_{\text{obs}}, \mathcal{C}, \mathcal{M}|N_l, N_r, L) \equiv x(T_{\text{obs}}, L) \Delta(T_{\text{obs}}|N_l, N_r) \approx F(T_{\text{obs}}|L) \Delta(T_{\text{obs}}|N_l, N_r), \quad (15)$$

where we implicitly define x and explicitly define

$$\Delta(T_{\text{obs}}|N_l, N_r) \equiv \int_0^{T_{\text{obs}}} d\chi_l^2 h(\chi_l^2|N_l) H(\chi_r^2|N_r) \Big|_{T_{\text{obs}}-\chi_l^2}^{T_{\text{obs}}} \quad (16)$$

and $H(\chi_r^2|N_r)$ is the cumulative distribution for the density h . H can be expressed in terms of the well known cumulative distribution function of χ^2 with k degrees of freedom, and requires no numerical integral. With the approximation

$$x(T_{\text{obs}}, L) = F(T_{\text{obs}}|L), \quad (17)$$

Equation (9) now simplifies to

$$F(T_{\text{obs}}|L) = F(T_{\text{obs}}|N_l)F(T_{\text{obs}}|N_r) - F(T_{\text{obs}}|L)\Delta(T_{\text{obs}}|N_l, N_r) \quad (18)$$

or

$$F(T_{\text{obs}}|L) = \frac{F(T_{\text{obs}}|N_l)F(T_{\text{obs}}|N_r)}{1 + \Delta(T_{\text{obs}}|N_l, N_r)}. \quad (19)$$

We expect this expression to become exact as $F(T_{\text{obs}}|L) \rightarrow 1$, and to show some discrepancies at smaller values of $F(T_{\text{obs}}|L)$ where the approximation employed in Eq. (17) is not valid. Since we underestimate the correction term, we overestimate $F(T_{\text{obs}}|L)$. The error is larger at values of T_{obs} where F is finite but not close to 1. We evaluate this effect for some examples below. For the more interesting region where $F \rightarrow 1$, we expect our approximation to be excellent.

The correction $\Delta(T_{\text{obs}}|N_l, N_r)$ is nearly independent of N_l, N_r in our region of interest (N_l and N_r large) due to the Bernoulli suppression of long runs in h which is $\propto 2^{-N_{l,r}}$; cf. Eq. (12). We provide numerical results supporting this claim in Sect. 4.

Let us consider the special case $N_l = N_r = N$ and assume that indeed $\Delta(T_{\text{obs}}|N, N)$ is independent of N , so

$$F(T_{\text{obs}}|2N) = \frac{F(T_{\text{obs}}|N)^2}{1 + \Delta(T_{\text{obs}})}. \quad (20)$$

Taking $n = 2$ as the base case, we can generalize to arbitrary $n \geq 2$ by induction to arrive at our main result

$$F(T_{\text{obs}}|nN) = \frac{F(T_{\text{obs}}|N)^n}{(1 + \Delta(T_{\text{obs}}))^{n-1}}, \quad n \geq 2. \quad (21)$$

To highlight the hidden assumptions in this procedure, we write down the induction step from $n \rightarrow n+1$ in detail using $N_l = nN, N_r = N$ and suppressing the T_{obs} dependencies

$$F(\cdot|(n+1)N) = F(\cdot|nN)F(\cdot|N) - x(\cdot, (n+1)N)\Delta \quad (22)$$

$$= F(\cdot|nN)F(\cdot|N) - F(\cdot|(n+1)N)\Delta \quad (23)$$

$$= \frac{F(\cdot|N)^{n+1}}{(1 + \Delta)^{n-1}} - F(\cdot|(n+1)N)\Delta \quad (24)$$

$$= \frac{F(\cdot|N)^{n+1}}{(1 + \Delta)^n}. \quad (25)$$

So the approximation $x(T_{\text{obs}}, nN) = F(T_{\text{obs}}|nN)$ is consistently employed $n-1$ times and we neglect contributions from runs longer than $2N$ which is acceptable for the same reasons that Δ can be considered independent of N . For concreteness, in our tests we set $\Delta \equiv \Delta(T_{\text{obs}}|N) \equiv \Delta(T_{\text{obs}}|N, N)$.

With this scaling equation, we can use exact results for moderate values of N to find the p value for our T_{obs} for very large $L = nN$.

3.2 Trials factor

In many applications, the look-elsewhere effect just amounts to multiplying a local p value for seeing a peak in one location by the total number of searched locations, or trials factor. Although this proportionality is usually not exact, it provides some insight into the inner workings of the look-elsewhere effect. Dividing the observations into n batches of size N and neglecting the denominator in Eq. (21), the overall p value

$$p(T_{\text{obs}}|nN) \approx 1 - (1 - p(T_{\text{obs}}|N))^n \approx n p(T_{\text{obs}}|N), \quad (26)$$

if both $p(T_{\text{obs}}|N)$ and Δ are small. This leads to the simple conclusion that the trials factor is just the number of batches

Fig. 3 Top: cumulative distribution, $F(T|L)$, for $L = 10^2, 10^3, 10^4, 10^5$. Bottom: probability density, $P(T|L)$, for the same L

n . We remark that with SQUARES, the trials factor is only approximately constant as it mildly depends on T_{obs} and possibly N . In the following, we do not use the trials factor because we have the exact relation given by Eq. (21) but we found the trials factor a useful tool for back-of-the-envelope estimates.

4 Numerical tests

We first display the behavior of $F(T|L)$ and $P(T|L)$ for different L in Fig. 3. For $L = 100$, the exact calculation is used, while for larger L the scaling Eq. (21) is used. As seen in the figure, the shape of the probability distribution does not change very much, but only shifts to larger values of T with a speed approximately proportional to $\log(L)$.

4.1 Δ Dependence on N

We now show that the quantity $\Delta(T|N)$ is indeed independent of N at large enough N . Figure 4 compares $\Delta(T|N)$ for several values of T . When T is large, $\Delta(T|N)$ is tiny and so is the p value. While $\Delta(T|N)$ indeed shows variations with N at small N , the dependence on N is negligible at larger N . For example for $T = 512$, $\Delta(T|N)$ is well into the saturated region at $N = 100$ and $\Delta(512|100, 100) = 10^{-84}$. Since we propose to use our approximation for $N \gtrsim 100$, the error introduced by ignoring the N dependence of Δ is completely negligible.

Fig. 4 $\Delta(T|N, N)$ as a function of N for several values of T

4.2 Accuracy of the approximation

We now investigate the size of the uncertainty introduced by our approximations in deriving our scaling formula. For this purpose, we compare the exact calculation for $F(T|L = 100)$ with the approximation for $N_l = N_r = 50$ in Fig. 5 (upper left plots). The top panel shows the difference in the cumulative distributions, while the lower panel shows the fractional difference in the p value as a function of the p value. As is seen in the figure, our scaling formula gives a very good agreement with the exact calculation with a maximum difference in the cumulative function of about 7×10^{-4} . It is also found that the difference between the exact calculation and the approximate calculation decreases as L and T increase. The fractional error on the p value at small p values is negligible.

To further study the accuracy of Eq. (21), we take the difference of the cumulative distributions, $F(T|n \cdot 100) - F(T|2n \cdot 50)$ for $n = 10, 100, 1000$ and plot these versus the value of the test statistic in Fig. 5. We see here that the calculations agree to better than the per mil level, with the largest differences visible for moderate values of the cumulative distribution function. This is exactly the region where the approximation used in Eq. (17) was expected to show small deviations. The correction term as evaluated is too small, so that $F(T|2n \cdot 50) > F(T|n \cdot 100)$. The effect decreases at larger T . In the most interesting case, where the p value is small ($F(T)$ large), the calculations are in excellent agreement with negligible differences.

Finally, we compare the calculated cumulative distribution with results from Monte Carlo simulations using the

MT19937 random-number generator [12] available in the Gnu Scientific Library [13]. The Box–Muller algorithm from the Gnu Scientific Library is then used to calculate Gaussian random numbers [14]. For the Monte Carlo simulations, we generate a large number of experiments with different-length runs, as specified in the table, and keep track of the value of the test statistic in each experiment. We then form the Monte Carlo estimate of the probability and compare to the calculated cumulative distribution function. We estimate the statistical uncertainty on the Monte Carlo result using the binomial probability standard deviation.¹ The results are shown in Fig. 6. For the analytic calculations, we used the $N = 100$ set of results.

For the case $L = 100$ we compare the simulation with the random number generator directly with the exact calculation as a test of the quality of the generator. This test is shown in the upper left panel of Fig. 6. As can be seen in the figure, the results are within the statistical fluctuations expected from the binomial distribution.

Given that we have a good random number generator, we can then check the agreement between our calculation given in Eq. (21) for different $n = 10, 100, 1000$ (see Table 1). The tests are shown in the other three panels. It is seen that in all cases the differences are in agreement with expected statistical fluctuations.

4.3 Precision of the calculation

A typical use of our results will be to evaluate the p value for an observed excess, where small p values will generate interest in follow-up analyses. The p value for the SQUARES statistic in a sequence of L measurements is given by

$$p = 1 - F(T|L). \quad (27)$$

Values of $F(T|N)$ for $N \leq 100$ can be calculated exactly to double floating-point precision using the results presented in [1, 11]. For $L > N$, we evaluate $F(T|L)$ by finding the value $n = L/N$ and evaluating $F(T|nN)$ using Eq. (21) even if $n \notin \mathbb{N}$. For the test case $L = 355$, we verified that the three alternatives ($n = 3.55, N = 100$), ($n = 5, N = 71$), and linear interpolation between ($n = 4, N = 88$) and ($n = 4, N = 89$) agree to nine significant digits.

The precision can be evaluated as

$$\begin{aligned} dF(T|nN) &= \frac{nF(T|N)^{n-1} dF(T|N)}{(1 + \Delta(T))^{n-1}} \oplus \frac{(1 - n)F(T|N)^n d\Delta(T)}{(1 + \Delta(T))^n} \end{aligned} \quad (28)$$

¹ Note that the resulting uncertainties are correlated since we are using the Monte Carlo estimates from the same set of experiments.

Fig. 5 First column: the difference between the exact calculation for $F(T|100)$ and the scaling equation 21 for $F(T|2 \cdot 50)$ is shown in the upper panel, while the lower panel shows the fractional error in the p value as a function of the p value. For the remaining plots, we compare $F(T|n \cdot 100)$ with $F(T|2n \cdot 50)$. Second column: $n = 10$. Third column:

$n = 100$. Last column: $n = 1000$. The value of n for the $N = 50$ case is always twice that used for $N = 100$ so that $L = Nn$ is the same. The value plotted in the upper plots is $F(T|n \cdot 100) - F(T|2n \cdot 50)$, the second plot shows the cumulative distribution, and the third plot the $p = 1 - F$ value on the log scale

Fig. 6 Comparison of Monte Carlo generated cumulative distributions with our calculations based on $N = 100$. First column: $n = 1$. Second column: $n = 10$. Third column: $n = 100$. Last column: $n = 1000$. The values plotted are $F(T|n \cdot 100) - F_{\text{Monte Carlo}}(T|n \cdot 100)$

$$\leq n \, dF(T|N) \oplus (1 - n) \, d\Delta(T), \quad (29)$$

where the inequality holds since $F(T|N) \leq 1$ and $\Delta(T) \geq 0$, and $\oplus$ indicates addition in quadrature. To reach a given precision ϵ on p , we require that $F(T|N)$ and $\Delta(T)$ be evaluated to an absolute precision ϵ/n . As an example, for $L = 10^6$, $N = 100$, and $\epsilon = 10^{-5}$, we would need an absolute precision on $F(T|N)$ and $\Delta(T)$ at the 10^{-9} level. We have verified that this can be reached in practice: regarding $F(T|N)$, our two implementations [11] in *mathematica* and *C++* agree at the 10^{-15} level, and with 1D numerical integration, $\Delta(T)$ can be computed at the 10^{-10} level.

5 Test case – (Fake) axion search

As an example of the use of our SQUARES statistic, we consider an experimental setup at the Max Planck Institute for Physics designed to search for axions in the mass range $40 - 400 \, \mu\text{eV}$ [10]. The measurement is effectively

Table 1 The Monte Carlo data sets generated. The length of the sequences and the numbers of simulated experiments are given. The last column gives the parameters used for the calculation

L	Experiments	Calculation
100	10^9	Exact analytic calculation
1000	10^7	$N = 100, n = 10$
10000	10^6	$N = 100, n = 100$
100000	10^6	$N = 100, n = 1000$

a power spectrum as a function of the frequency of emitted microwave radiation built from many independent measurements. The baseline signal is dominated by the thermal background (10^{-19} W). A weak fake axion signal was injected; the location and width of the signal were unknown when the analysis was carried out. Although the shape (Gaussian) was known, this information was not used. The spectrum to be analyzed is shown in Fig. 7. The shape of the background spectrum and level of fluctuation was unknown, and had to be determined from the data by assuming that any possi-

Fig. 7 The power spectrum to be analyzed. The data are from a test setup for an axion search experiment developed at the Max Planck Institute for Physics. A signal corresponding to a possible axion signal was introduced in the setup and searched for using the test statistic described in this paper. The arrow shows the location where the signal was injected. The region of the signal is shown in the inset together with the background fit (top inset), the residuals from the fit (middle inset) and the run χ^2 (lower inset)

ble signal would provide a much sharper feature than any change in background. In the example considered here, the spectrum consists of 24,576 data points giving the integrated power in ≈ 2 kHz intervals. A signal is expected to be one or a few bins wide. To minimize the number of assumptions made about the signal shape, we scanned the spectrum with SQUARES to determine if there was a significant deviation from background expectations.

To determine the background shape and fluctuations, the full spectrum was partitioned into contiguous subsets and a second-order polynomial fit was used to find the background shape. The residuals were used to extract the standard deviation of the fluctuations. Different lengths of the subsets were considered from a minimum of 96 measurements to 256 measurements. It was verified that the residuals followed the expected Gaussian distribution.

The distribution of T for the 256 sets of 96 data points is shown in Fig. 8, and compared to the expectation from the exact calculation of [1] for $L = 96$. The largest value found was $T_{\text{obs}} = 57.3$. For $L = 96$, that is if the total number of observations had been only 96, this has a p value of $p = 6.4 \times 10^{-9}$. To get the p value for all observations, we use Eq. (21) and find a p value of $p = 1.9 \times 10^{-6}$, which is very small. The frequency at which this signal was found was indeed the frequency of the injected 'fake axion'. For comparison, the second most significant p value of a test

Fig. 8 The distribution of the test statistic calculated in the 256 sets of 96 samples, compared to the expected distribution for $L = 96$

statistic found in one of our runs was 2×10^{-5} taking $L = 96$, which becomes $p \approx 6 \times 10^{-3}$ when taking the full 24576 samples into account. These changes of the p value illustrate the importance of the look-elsewhere effect. To infer the p value from N for nN observations, the obvious guess for the trials factor in Eq. (26) is 256, which is too small by just 20% for the above numbers. Our more accurate result based on Eq. (21) is achieved for essentially the same computational effort.

Figure 7 shows in the inset the data in the range of the candidate signal together with the polynomial background fit, the residuals, and the value of S as a function of the frequency. The test was repeated several times with different fake signals and these were found in every case.

To appreciate the usefulness of Eq. (21), we consider the computational cost. Proposition 1 from [1] states that computing the exact expression for the distribution of T for $L = 24576$ requires a sum over the enormous number of 2.6×10^{169} integer partitions, something that is completely unfeasible with the best supercomputers today. But with our approximate result, we only need to compute the exact expression for $N = 96$ which requires 1.4×10^8 partitions, or a reduction of work by 161 orders of magnitude. The algorithm is of the streaming type and the partitions need not be stored but can be independently processed, so it makes ideal use of modern multi-core computers. On a desktop computer with an Intel i7-4700 CPU with four cores and eight openMP threads, the C++ implementation [11] of $F(T|N = 100)$ is computed in 1.8 s. With some reduction in precision and range of validity, computing for $N = 50$ as the baseline requires less than 10 ms.

6 Conclusion

We presented an extension of our previous work [1] where we introduced the SQUARES statistic, T , to detect bumps that are incompatible with a background model for ordered 1D data sets such as spectra. We derived the exact distribution of the statistic needed to compute a p value before but the expression could not be evaluated beyond about 100 data points within reasonable time. In this work, we describe an approximation Eq. (21) that takes the results from few data points and extrapolates to millions of points in a principled manner by dividing the data into subsets. In the region of interest where the observed value T_{obs} is large and the p value is small, the approximation has both high accuracy and precision. We recommend to use the exact expression for 100 data points as the basis for extrapolation but in most applications one may even start lower without loss of precision. The largest discrepancies between exact and approximate distribution appear for intermediate values of T but even there they are too small to change the judgment of the quality of the model. The code implementing all expressions in this paper is available online at <https://github.com/fredRos/runs>.

Through Monte Carlo experiments with up to 10^5 data points we validated our analytic result. In addition, the test statistic is computed for a real-life physics problem of detecting a fake axion signal in a power spectrum of 24,576 data points and shows a very significant excess from the background model at the expected location. All other properties studied in this example are in full agreement with our derivations. This provides the basis to apply this method when the axion experiment has grown to full scale with millions of data points.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. Funded by SCOAP³.

References

1. F. Beaujean, A. Caldwell, J. Stat. Plann. Inference **141**(11), 3437 (2011). <https://doi.org/10.1016/j.jspi.2011.04.022>
2. F. James, *Statistical methods in experimental physics* (World Scientific, Singapore, 2006)
3. N.A. Heard, P. Rubin-Delanchy, Biometrika **105**(1), 239 (2018). <https://doi.org/10.1093/biomet/asx076>
4. F. Beaujean, A. Caldwell, D. Kollár, K. Kröninger, Phys. Rev. D **83**, 012004 (2011). <https://doi.org/10.1103/PhysRevD.83.012004>
5. G. Aad et al., Phys. Lett. B **716**, 1 (2012). <https://doi.org/10.1016/j.physletb.2012.08.020>
6. S. Chatrchyan et al., Phys. Lett. B **716**, 30 (2012). <https://doi.org/10.1016/j.physletb.2012.08.021>
7. G. Cowan, K. Cranmer, E. Gross, O. Vitells, Eur. Phys. J. C **71**, 1554 (2011). <https://doi.org/10.1140/epjc/s10052-011-1554-0>, <https://doi.org/10.1140/epjc/s10052-013-2501-z>. [Erratum: Eur. Phys. J. C **73**, 2501 (2013)]
8. O. Vitells, In: Proceedings, PHYSTAT 2011 workshop on statistical issues related to discovery claims in search experiments and unfolding, CERN, Geneva, Switzerland 17–20 January 2011. CERN (CERN, Geneva, 2011), pp. 183–189. <https://doi.org/10.5170/CERN-2011-006.183>
9. E. Gross, O. Vitells, Eur. Phys. J. C **70**, 525 (2010). <https://doi.org/10.1140/epjc/s10052-010-1470-8>
10. A. Caldwell, G. Dvali, B. Majorovits, A. Millar, G. Raffelt, J. Redondo, O. Reimann, F. Simon, F. Steffen, Phys. Rev. Lett. **118**(9), 091801 (2017). <https://doi.org/10.1103/PhysRevLett.118.091801>
11. F. Beaujean (2018). <https://doi.org/10.5281/zenodo.1323369>
12. M. Matsumoto, T. Nishimura, ACM Trans. Model. Comput. Simul. (TOMACS) **8**(1), 3 (1998)
13. B. Gough, *GNU scientific library reference manual* (Network Theory Ltd., United Kingdom 2009)
14. G.E. Box, M.E. Muller, Ann. Math. Stat. **29**(2), 610 (1958)