The Squares Statistic for Bump-Hunting in MADMAX

AC

May 10, 2025

1 Introduction

This is a simple first study of bump-hunting in the MADMAX data where we study the interplay between the width of a signal and the binning used to perform the analysis. We investigate the use of the Squares statistic for this purpose as this statistic does not require the definition of a line shape. This makes the search more general, allowing for the recognition of narrow or wide signals, but it is expected that for a predefined line shape a search strategy making use of the line shape will be more sensitive. Both types of search strategies should be used. In this note, we look at how the recognition of a signal depends on the amplitude and width of the signal in relation to the binning chosen for the analysis.

2 Review of the Squares statistic

The Squares statistic is part of the RunStatistics.jl package. It is a test statistic based on runs of deviations from a baseline expectation and assumes Gaussian distributed fluctuations of the data. It supplements the classic χ^2 test which ignores the ordering of observations and provides sensitivity to local deviations from expectations. The exact distribution of the statistic in the non-parametric case is derived and an algorithm to compute p-values is presented in [1]. Its application to long sequences of data, as will be the case for MADMAX, is described in [2]. The package RunStatistics.jl is an implementation of the code to carry out the calculations.

Specifically we assume there are L observations y_i and the index i provides an ordering for the data. We assume the observations independently follow a Gaussian or Normal distribution

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

where the expectation μ and the standard deviation σ are known for every i. In the MADMAX analyses to date, the μ_i are determined using a Savitzky-Golay Filter [3] as the shape of the background is not know. The σ_i are also determined in the background fitting procedure and are assumed known. For the analyses described in this note, we assume that the baseline has been subtracted so that $\mu_i = 0 \ \forall\, i$ and also that the amplitudes are normalized such that $\sigma_i = 1 \ \forall\, i$

The main motivation behind the runs statistic is that it is signal shape agnostic and it automatically takes care of the look-elsewhere effect (also called the trials factor) that arises in some other methods. These typically look for narrow peaks where typically the profile-likelihood ratio statistic was employed [4] which requires fully specifying both the background and the signal model including dependence on unknown parameters to be estimated from the data.

In comparison, the runs statistic does not require a signal model and does not rely on asymptotic normality of the likelihood but assumes the background is known exactly. In [1], it was demonstrated that in this setting the runs statistic leads to a more powerful test than the classic χ^2 test in this peak-fitting problem.

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, $y_i \geq \mu_i$. The runs statistic T is defined as the largest value of χ^2 for any success run

$$T \equiv \max_R \sum_{i \in R} \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2 , \tag{2}$$

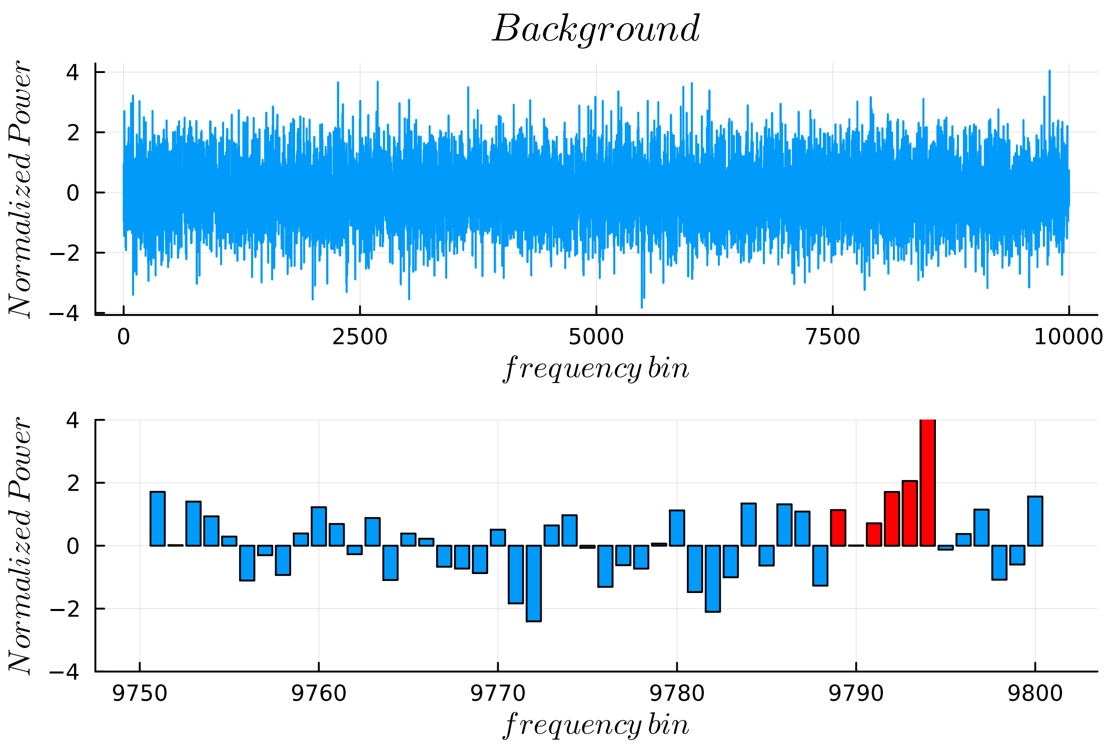

Figure 1: Top plot: A background data set. Bottom plot: The zoom of the data set shown in the top plot. The run generating T_{obs} is highlighted in red.

where R represents the set of indices in an individual success run. Using the cumulative $F(T|L)$, the p value is the tail-area probability to find T larger than the observed value T_{obs},

$$p \equiv 1 - F(T_{\text{obs}}|L)\,. \tag{3}$$

Note that χ_i^2 is 0 initially and as soon as a failure is encountered; i.e., $y_i < \mu_i$. Otherwise, it is incremented by $(y_i - \mu_i)^2/\sigma_i^2$ for every success. The exact probability distribution for the runs statistic T has been derived in [1]. Its use for $L \gtrapprox 80$ and the resulting approximate formulas are given in [2].

3 Setting up the calculations

In this initial study, we assume as a starting point that we have a set of 10000 power measurements centered on 0 and fluctuating according to the unit normal distribution. This will be the background-only case. As an example, we display one data set in the top plot of Fig. 1. The data shown is just 10000 samples from the unit normal distribution.

For this particular random sample, $T_{\text{obs}} = 25.4$ and occurs for the run from bin 9789-9794. The blowup of this region is shown in the lower plot of Fig. 1. The p value calculated for this T_{obs} - $p = 0.116$, so we would probably not get excited by this particular run. There is an 11.6 % chance to get such a p-value or lower in any random data set of this length.

As a sanity check that our p-value calculation is as expected, we can see what the p-value distribution looks like for a large sample of background runs where no signal is present. If our p-value calculation is correct, we expect a flat distribution between $p = 0$ and $p = 1$. The result is shown in the top plot of Fig. 2 and gives the expected result.

We introduce signals in the following way:

1. Signals are simulated as Gaussian distributions centered in our 10000 bin power distribution. One can pick different line shapes in the future to see how strongly the results vary with different choices. We use a Gauss distribution just for simplicity in this initial analysis.
2. The amplitude, A, and width, σ_{sig}, of the signal are varied as indicated in the tables below. The amplitude is defined relative to the background noise level. E.g., an amplitude of 10 means that the integrated signal

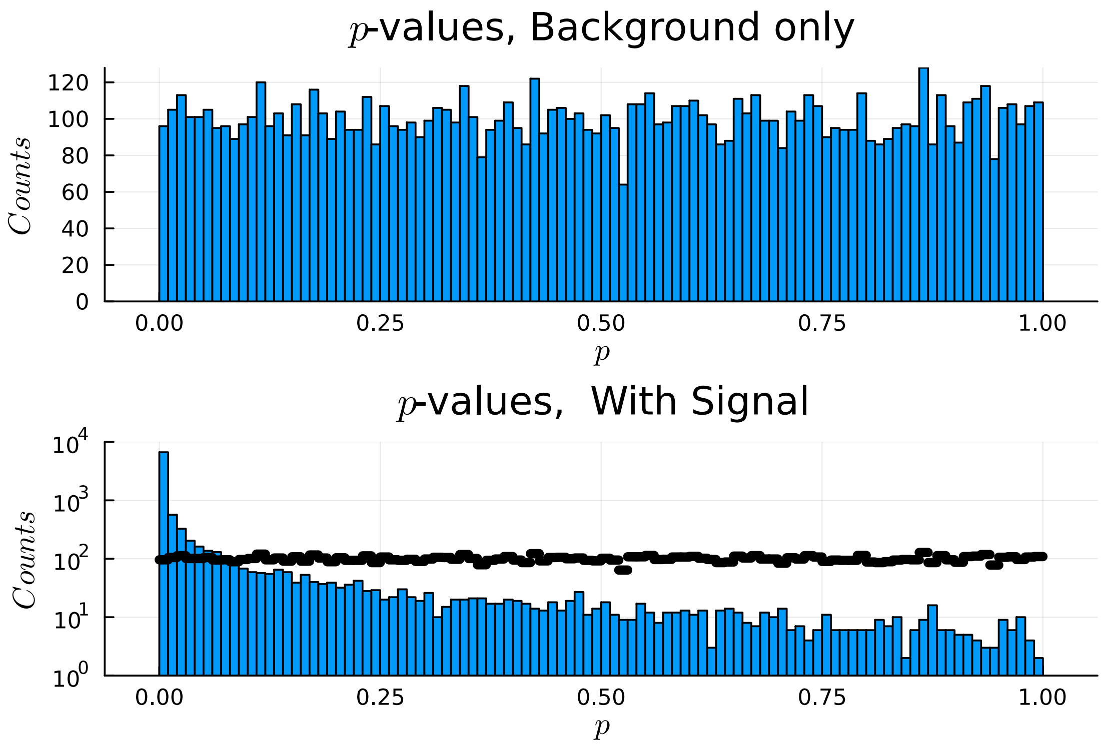

Figure 2: Top plot: the p-value distribution for the T_{obs} statistic for a large number of background only data sets. Bottom plot: the p-value distribution for the T_{obs} statistic for a large number of background + signal data sets, where the signal is sampled from a Gaussian with amplitude 20 (in units of the background noise) and the width is 4 bins. Note the log scale on the lower plot. The background only distribution is shown in black in the lower plot.

strength over all bins is ten times the background fluctuation level. The width is defined in terms of the (finest) bin width of the spectrum. A width of 4 means that the Gaussian standard deviation of the signal is four times the finest bin size.

3. The power in the signal is assumed to always be much smaller than the noise power, so that the fluctuations of background+signal are the same as the fluctuations of background only. I.e., in our setup, with unit normal fluctuations but with the mean in a bin adjusted to account for the presence of a signal.

The code for performing these tests is shown below. An example of what happens to the p-value distribution with $A = 20$ and $\sigma_{\mathrm{sig}} = 4$ is given in Fig. 2, where we have performed the analysis with the finest binning. The p-value peaks sharply at 0 in this case (note the log scale) indicating that we have an excellent chance to recognize the presence of the signal. If we make, e.g., a cut requiring $p \leq 0.05$, then 5 % of the time we will have a positive outcome in the background only case, while retaining > 99 % of the signal cases.

To test the effect of the interplay between the bin widths and the signal width, 4 different bin widths were chosen. These are the initial bin width, and groupings of (2,5,10) bins at a time. The expectation being that if the bins are too fine, it will be hard to see a wide signal due to the background fluctuations, whereas if the bins are too coarse, it will be hard to see a narrow signal.

4 Results

The p-values distribution for a series of 8 different amplitudes, A, and 6 different signal widths, σ_{sig} was extracted. The p-values distributions were evaluated for the finest bin width, as well as groupings of (2,5,10) bins. In the two tables below, the mean of the p-values are given (Table 1) and the fraction of trials with $p \leq 0.05$ (Table 2).

As seen in Table 1, the average p-value is close to 0.5 for the smallest signal amplitudes, and starts to significantly become smaller as the signal amplitude reaches 10. This is also reflected in Table 2 where we see that the relative frequency to pass the $p > 0.05$ cut only starts to increase significantly as the signal amplitude reaches 10. For the

A	BinSize	$\sigma_{\rm sig}$ 2.0	$\sigma_{\rm sig}$ 4.0	$\sigma_{\rm sig}$ 6.0	$\sigma_{\rm sig}$ 8.0	$\sigma_{\rm sig}$ 10.0	$\sigma_{\rm sig}$ 12.0
0	1	0.5	0.49	0.49	0.5	0.5	0.53
0	2	0.5	0.49	0.51	0.5	0.5	0.52
0	5	0.5	0.5	0.5	0.5	0.51	0.52
0	10	0.51	0.49	0.49	0.5	0.5	0.5
4	1	0.51	0.48	0.49	0.49	0.48	0.5
4	2	0.49	0.49	0.5	0.49	0.49	0.5
4	5	0.5	0.49	0.5	0.49	0.49	0.5
4	10	0.49	0.49	0.49	0.49	0.48	0.48
8	1	0.4	0.46	0.49	0.47	0.49	0.5
8	2	0.4	0.45	0.47	0.47	0.49	0.49
8	5	0.44	0.47	0.48	0.49	0.49	0.48
8	10	0.47	0.47	0.48	0.48	0.49	0.49
12	1	0.18	0.39	0.45	0.46	0.47	0.5
12	2	0.19	0.36	0.43	0.45	0.47	0.48
12	5	0.29	0.37	0.41	0.44	0.45	0.47
12	10	0.4	0.42	0.42	0.44	0.45	0.46
16	1	0.024	0.21	0.36	0.42	0.46	0.46
16	2	0.029	0.18	0.29	0.37	0.4	0.43
16	5	0.096	0.21	0.29	0.35	0.38	0.39
16	10	0.27	0.27	0.31	0.34	0.38	0.38
20	1	0.00068	0.072	0.22	0.34	0.41	0.44
20	2	0.001	0.052	0.15	0.25	0.34	0.38
20	5	0.011	0.062	0.15	0.21	0.29	0.32
20	10	0.15	0.14	0.18	0.22	0.29	0.33
24	1	1.0e-6	0.014	0.1	0.21	0.34	0.38
24	2	7.4e-6	0.0084	0.046	0.13	0.23	0.31
24	5	0.00049	0.013	0.044	0.098	0.17	0.22
24	10	0.038	0.043	0.062	0.11	0.17	0.22
28	1	7.1e-10	0.00058	0.029	0.12	0.26	0.3
28	2	1.7e-9	0.00043	0.011	0.052	0.13	0.2
28	5	1.3e-6	0.0006	0.011	0.038	0.086	0.13
28	10	0.0051	0.0078	0.018	0.047	0.092	0.12

Table 1: Mean of p-values. The first column gives the amplitude,A, of the input signal. The second column give the width of the bins in which power is measured, and the succeeding columns give the width of the input signal $\sigma_{\rm sig}$.

smallest value of $\sigma_{\rm sig}$ investigated, the percentage of trials passing the selection cut crosses 90 % around signal strength $A = 16$ for the finest power bin widths. As the binning gets coarser, the chance to recognize the signal decreases as expected.

As the width of the signal increases, the chance to identify the signal decreases for all bin sizes. This is not surprising, as we integrate more noise with increasing signal width. However, we see that the chance to recognize the signal increases with bin width for wide signal. E.g., looking at $A = 24$ and $\sigma_{\rm sig} = 10$, we see that the optimum bin width is 5 times the minimum bin width. This indicates that the optimum bin width to search for a signal depends on the width of the signal itself.

5 Discussion

The signal strength expected is a combination of several factors: the density, possibly time varying, of the source, its coupling strength to the photons in the magnetic field, and probably other factors. We should therefore think about search strategies where we want to reach a predefined sensitivity level for the 'plain' QCD axion with an assumption on its density (enough to saturate the expected DM density?) and its line shape, but also allow for unexpected signals at some level. These chosen sensitivity levels, defined e.g. as a frequentist Type-2 error rate for the different types of signals searched for, can then be converted into p-value cuts for the Squares statistic and possibly other test statistics that we want to use. This will be an interesting discussion!

A	BinSize	σ_{sig} 2.0	σ_{sig} 4.0	σ_{sig} 6.0	σ_{sig} 8.0	σ_{sig} 10.0	σ_{sig} 12.0
0	1	0.043	0.048	0.039	0.051	0.039	0.041
0	2	0.045	0.052	0.039	0.038	0.05	0.034
0	5	0.064	0.041	0.043	0.057	0.054	0.053
0	10	0.046	0.059	0.046	0.046	0.056	0.049
4	1	0.04	0.05	0.061	0.052	0.059	0.052
4	2	0.044	0.056	0.05	0.045	0.058	0.058
4	5	0.044	0.049	0.054	0.049	0.062	0.057
4	10	0.045	0.06	0.052	0.055	0.062	0.043
8	1	0.13	0.075	0.05	0.057	0.06	0.04
8	2	0.13	0.081	0.072	0.067	0.048	0.053
8	5	0.088	0.066	0.064	0.052	0.05	0.06
8	10	0.079	0.04	0.072	0.059	0.05	0.05
12	1	0.52	0.15	0.09	0.06	0.065	0.057
12	2	0.48	0.17	0.11	0.073	0.071	0.065
12	5	0.28	0.16	0.13	0.084	0.084	0.084
12	10	0.13	0.11	0.11	0.081	0.083	0.065
16	1	0.9	0.45	0.21	0.13	0.084	0.075
16	2	0.89	0.5	0.29	0.18	0.14	0.11
16	5	0.69	0.43	0.27	0.19	0.14	0.13
16	10	0.3	0.28	0.24	0.2	0.13	0.14
20	1	1.0	0.78	0.43	0.23	0.15	0.084
20	2	1.0	0.83	0.55	0.38	0.23	0.16
20	5	0.95	0.77	0.54	0.43	0.28	0.21
20	10	0.55	0.56	0.46	0.39	0.27	0.19
24	1	1.0	0.96	0.71	0.42	0.23	0.18
24	2	1.0	0.97	0.83	0.63	0.4	0.3
24	5	1.0	0.95	0.83	0.67	0.5	0.4
24	10	0.84	0.82	0.77	0.64	0.46	0.38
28	1	1.0	1.0	0.9	0.64	0.38	0.27
28	2	1.0	1.0	0.96	0.84	0.61	0.45
28	5	1.0	1.0	0.96	0.86	0.71	0.61
28	10	0.98	0.97	0.92	0.83	0.68	0.62

Table 2: The fraction of trials passing the $p \leq 0.05$ cut. The first column gives the amplitude,A, of the input signal. The second column give the width of the bins in which power is measured, and the succeeding columns give the width of the input signal σ_{sig} .

Some Code

```
#
# As a sanity check, we see if the p-value distribution is flat for background only
#
Random.seed!(1234);

nsanity=10000
pv=zeros(nsanity)
for i=1:nsanity
    Run=randn(nbins)
    Squares=t_obs(Run, 0., 1.)
    pv[i] = squares_pvalue_approx(Squares[1],Ns1, epsp)
end

# Now add a substantial signal and see the distorted p-value distribution

Run1=zeros(nbins)
A=20
w=4
ds = Normal(0.,w)
pvs=zeros(nsanity)

for i=1:nsanity
    for bin=1:nbins
        x1=(bin-1) - nbins/2
        x2=bin - nbins/2
        bin_integral=A*(cdf(ds,x2)-cdf(ds,x1))
#
        d = Normal(bin_integral,1.)
        Run1[bin]=rand(d)
    end
    Squares=t_obs(Run1, 0., 1.)
    pvs[i] = squares_pvalue_approx(Squares[1],Ns1, epsp)
end
```

References

[1] F. Beaujean and A. Caldwell, Journal of Statistical Planning and Inference 141, 3437 (2011).

[2] F. Beaujean, A. Caldwell and O. Reimann, Eur. Phys. J. C **78**, no.9, 793 (2018) doi:10.1140/epjc/s10052-018-6217-y [arXiv:1710.06642 [hep-ex]].

[3] A. Savitzky and M. J. E. Golay, Anal. Chem. **36**, no.8, 1627-1639 (2002) doi:10.1021/ac60214a047

[4] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C **71**, 1554 (2011) [erratum: Eur. Phys. J. C **73**, 2501 (2013)] doi:10.1140/epjc/s10052-011-1554-0 [arXiv:1007.1727 [physics.data-an]].