Analysis walk-through: Search for a Higgs boson in the 4 muon final state.

H. Alida F. Hardersen
Fysisk Institutt, Universitetet i Oslo

(Dated: September 1, 2020)

A statistical procedure to search for an excess in data corresponding to a higgs boson decaying to four leptons $H \to 4\mu$ using common methods for hypothesis testing is presented and applied to fake data. The analysis was not able to discover/exclude a 125 GeV Higgs boson with the 5σ confidence.

I. INTRODUCTION

In this analysis we will search for a signal on top of a Standard Model background. The histograms used includes 1. a Monte Carlo simulation of the expected signal from a 125 GeV Higgs boson decaying to four leptons, 2. MC simulation of the irreducible Z background, and 3. a set of "real" data. We will be using counting techniques to quantify the sensitivity of the search and find the optimal signal region. Then we will use a data-driven sideband fit to estimate the background contribution in the signal region. By using common statistical methods for hypothesis testing, we will try to determine weather there is an excess in the data, and if this excess is compatible with a 125 GeV Higgs boson.

The analysis is based on the chapter "Analysis Walk-through" in the book Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods[2]. The code is a continuation of the skeleton code written by Ivo van Vulpen and Aart Heijboer in 2013. The full code used in the analysis can be found in https://github.com/Alidafh/FYSSP100

II. THE ANALYSIS

The analysis begins by defining the hypothesis that is being tested. The null hypothesis H_0, which is the hypothesis that is expected to be true, is that the events are purely known SM-processes and no signal is present(background/SM hypothesis). The alternative hypothesis H_1 is that additional new physics processes contribute(signal+background hypothesis).

$$\begin{aligned} H_0 &: \text{ background-only} \\ H_1 &: \text{ signal+background} \end{aligned}$$

The next step is to define a quantity, t, that only depends on the data. This is called a test-statistic, and is used to determine the level of agreement between the data and the hypothesis. In the first part of the analysis, which is treated as a simple counting experiment, the test-statistic used will be the number of observed events, $t = n_{obs}$. Later the test-statistic used will be the profile likelihood ratio which will be defined later in this paper.

A. Significance and p-value

To claim discovery/exclusion of a signal we need to quantify the inconsistency between the data and the null-hypothesis, which can be done using the p_0-value and the significance. The p_0-value is a uniform distribution between 0 and 1, and a small p_0-value is an indication that the null-hypothesis does not give a good-enough description of the data.

In the counting experiment, where the observed number of events is used as the test-statistic, the p_0-value is the probability to count a number of events that is equal to or lager than the one we observed[1],

$$p_0 = P(t_{obs} \geq t) = \int_{t_{obs}}^{\infty} f(t; H_0)dt \tag{1}$$

where $f(t; H_0)$ is the probability density function(pdf), which in this analysis is the Poisson distribution described in [2, p 8]. Equation 1 can be rewritten as,

$$p_0 = \sum_{t=t_{obs}}^{\infty} f(t; H_0) = 1 - \sum_{t=0}^{t_{obs}} f(t; H_0) \tag{2}$$

To find this analysis's ability to separate the two hypotheses, the expected p-value and significance is used. This is calculated by using the median number of expected events under the alternative hypothesis H_1 as the test-statistic in equation 2, instead of the observed number of events in the data. The significance Z_0 is calculated from p-value using equation 3.[2]

$$Z_0 = \Phi^{-1}(1 - p_0) \tag{3}$$

where Φ^{-1} is the cumulative distribution function for the unit Gaussian. [2, p 84]

B. Data-driven background estimation

In real life we do not know the background level with absolute precision so in order to reduce the systematic uncertainties and receive the same efficiency in MC as we have in data we can do a fit to the data. Assuming the shape of the MC-simulations can be trusted, we can use the sideband region in the data to determine an estimate, α, that is used to normalise/reweigh the MC.

The sideband is defined as the region where we do not expect to see any signal. We can parametrize the combined signal+background distribution as a function of 4-lepton invariant mass,

$$f_{total} = \mu s + \alpha b \tag{4}$$

where s and b are the expected distributions of events for the signal and background respectively. We can see that the null hypothesis H_0 is simply a special case of H_1 with $\mu = 0$.

1. The method of maximum Likelihood

One of the most frequently used methods for parameter estimation is the method of maximum likelihood, where the best estimates are those that that maximise the likelihood function. These estimates are referred to as the maximum likelihood estimates (MLE). For an experiment with independent and identically distributed data, the likelihood function L is defined as the joint pdf for the whole data sample,

$$L(n; \mu, \alpha) = \prod_{i=1}^{N} f(n_i | \alpha b_i + \mu s_i) \tag{5}$$

here n_i is the observed number of events in bin i, and s_i and b_i are the expected number of events in the ith bin for signal and background respectively. For the sideband fit, the signal scale-factor is set to zero, $\mu = 0$. Because sums are nicer to work with than products, is often more convenient to minimise the negative log-likelihood function instead,

$$-\ln L(n; \mu, \alpha) = -\sum_{i=1}^{N} \ln f(n_i | \alpha b_i + \mu s_i) \tag{6}$$

where f is the Poisson distribution. To determine the 1σ uncertainties on the found ML estimate $\hat{\alpha}$, we can do a scan of the log-likelihood around its maximum value,

$$\Delta \ln L = \ln L - \ln L_{max} = -1/2 \tag{7}$$

$$-2 \ln L = -2 \ln L_{max} + 1 \tag{8}$$

The two points found will correspond to the interval $[\Delta\hat{\alpha}_-, \Delta\hat{\alpha}_+]$ that contains, in 67.27% of cases, the true value of α.

C. Likelihood ratio test-statistic

As mentioned earlier, we will also use a more complex test statistic. This test statistic originates in the Neyman-Pearson lemma which states that for that in the case of simple hypothesis, the most optimal choice of test-statistic is the Likelihood ratio,

$$Q = \frac{L(\mu = 1)}{L(\mu = 0)} \tag{9}$$

where $L(\mu = 1)$ is the likelihood function assuming H_1 and $L(\mu = 1)$ is the likelihood assuming H_0. When there are a large number of measurements available, Wilk's theorem allows finding an approximate asymptotic expression for a test-statistic that is based on this likelihood ratio,

$$t = -2 \ln Q \tag{10}$$

We can generate multiple pseudo-datasets for each hypothesis for which we calculate this test-statistic in order to get the test-statistic pdf. These will be denoted as $g(t|H_0)$ for the null hypothesis, and $g(t|H_1)$ for the alternative hypothesis.

1. Rules for discovery/exclusion

We need to set some rules as to what ranges of values for the test statistic will lead to discovery/exclusion. We want to say something about the significance of the discovery/exclusion using a confidence level, which is the compatibility of the test statistic observed from data t_{obs} with the given hypothesis. The confidence in the null hypothesis, H_0, is given by the probability that the test-statistic is less than, or equal to the observed value from the data,

$$CL_b = P(t \leq t_{obs} | H_0) = \int_{-\infty}^{t_{obs}} g(t|H_0) dt \tag{11}$$

Values of CL_b close to 1 indicate poor compatibility with the null hypothesis and the alternative hypothesis is favoured. To claim discovery of a signal we need to reject this background hypothesis, the convention in high energy physics is that the p-value $1 - CL_b$ must be smaller than 5.73×10^{-7} to claim a 5σ discovery. The confidence in the alternative hypothesis, H_1 is given by the probability that the test statistic is equal to or larger than the observed value in the data,

$$CL_{sb} = P(t \geq t_{obs} | H_1) = \int_{t_{obs}}^{+\infty} g(t|H_1) dt \tag{12}$$

Values of CL_{sb} that are small indicate a non-compatibility with the alternative hypothesis and the null hypothesis is favoured, i.e. we can exclude a signal. One possibility is to say that the signal is excluded if $CL_{sb} < 5\%$, but this causes some issues.

2. The modified frequentist method - CL_s

When using the likelihood ratio test-statistic to calculate the pdf's for the two hypotheses, the two will overlap if the expected signal is low and if the number of observed events are below the expected background, none of the two hypotheses are favoured. Because of the low p-value for the signal-hypothesis, H_1, we may end up

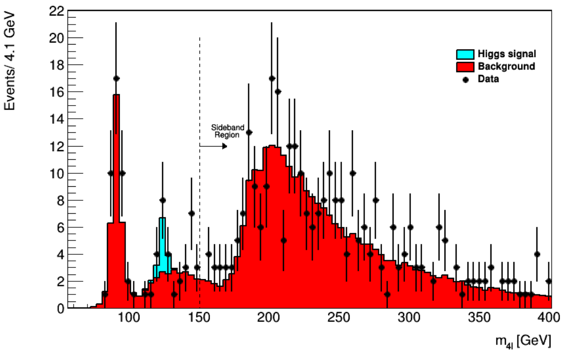


FIG. 1: The data files used in this project. Black dots are "real" data, red is background simulations and the blue is the signal simulations for a Higgs boson with mass $m_H = 125$ GeV

claiming that is excluded at high confidence, suggesting that we should move on to look at a different mass Higgs. A way to deal with this is to use the modified frequentist re-normalisation - CL_s method, where the confidence level for the signal+background hypothesis(alternative-hypothesis), CL_{sb} is normalized to the confidence level of the background hypothesis, CL_b,

$$CL_s = \frac{CL_{sb}}{CL_b} \tag{13}$$

This can be interpreted as an approximation to the confidence in the signal hypothesis one might obtain if the experiment had been preformed in absence of background, but it should be noted that it is not an actual confidence level.[3] The signal hypothesis will be excluded at confidence level CL when

$$1 - CL_s \leq CL \tag{14}$$

so we can exclude the signal hypothesis at 95% confidence if $CL_s < 5\%$. If the alternative signal+bsckground hypothesis is well separated from the background hypothesis, $CL_s \approx CL_{sb}$.[4] [5]

III. RESULTS

A. Finding the optimal mass window for the signal region

If a Higgs boson of $m_H = 125$GeV is present in the data it is expected to present as an excess of events around $m_{4l} = 125$GeV. Therefore, we need to determine a window around this mass that optimises the expected significance described by equation 3 so that we have the best compromise between the signal efficiency and the background rejection.Figure 2 shows the significance as a function of signal region width for both the expected and observed case. As can be seen in the figure, the observed significance is at a maximum when the mass window is small. The expected significance has a maximum of 2.04σ at a mass window of 7.15GeV.Figure 3 shows the expected significance for 5 times higher luminosity. Here, there is a maximum significance of 4.79σ for a mass window of 6.55 GeV. The lowest possible luminosity scale factor needed to make a discovery is 5.51, where the expected significance will be 5.07σ when using a mass window of 6.15 GeV.

TABLE I: The maximum expected significance and corresponding mass window around 125 GeV for various luminosities

Luminosity scale	$Z_{exp}[\sigma]$	Mass Window [GeV]
1	2.04	7.15
5	4.79	6.55
5.51	5.07	6.15

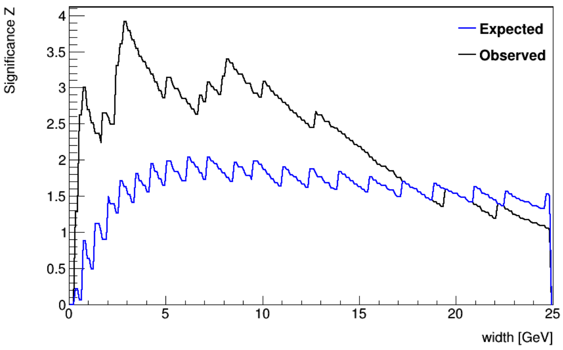


FIG. 2: The expected and observed significance as a function of mass window around 125 GeV.

B. Background estimation from sideband

To find the contribution of the background in the signal region, the scale factor α is determined using a likelihood

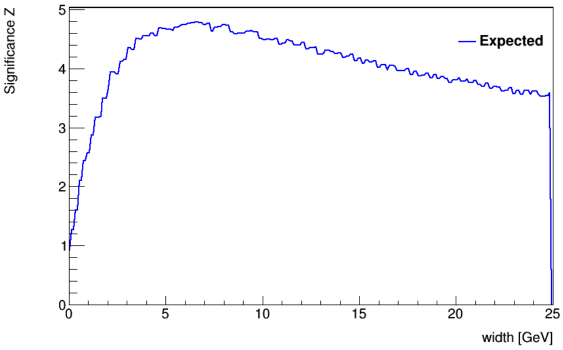


FIG. 3: The expected significance as a function of mass window around 125 GeV for 5 times higher luminosity.

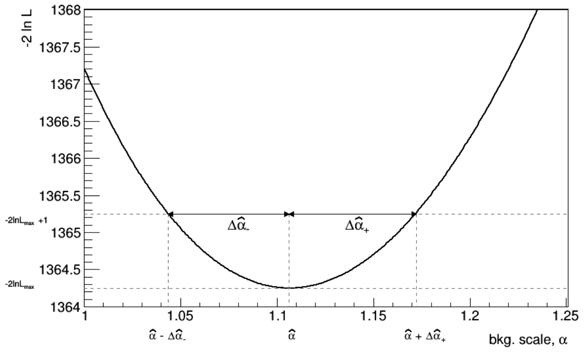


FIG. 4: The negative log likelihood function as a function of background scale factor α. The best fitted value $\hat{\alpha}$ correspnds to the minimum of the likelihood function. Also marked in the figure is the $\Delta\hat{\alpha}_{\pm}$ errors.

fit to the data in the sideband region which is defined as the range 150 GeV $\leq m_{4l} \leq$ 400 GeV as illustrated in figure 1. The likelihood function described by equation 6 is used with the signal scale factor set to zero ($\mu = 0$). A plot of the negative log-likelihood function $-2\ln L(\alpha)$ vs. background scale factor α is shown in figure 4. The likelihood function has the shape of a parabola, with a minimum which corresponds to the value of the best fitted estimator for the background, $\hat{\alpha}$. This is found to be close to one which indicates that the MC-simulations and the data are in good agreement.

$$\hat{\alpha} = 1.11^{+0.07}_{-0.06} \tag{15}$$

Figure 5 shows the comparison between the unscaled and scaled bacground MC samples and the data. Using the optimal mass window for luminosity scale 1 listed in table I, we find the scaled and unscaled number of background events in the signal region,

$$b = 4.64 \quad \rightarrow \quad b = 5.13^{+0.30}_{-0.29} \tag{16}$$

Using this new background estimate the expected significance is calculated by generating a set of toy experiments. For each toy experiment, a random Gaussian number is drawn for the background distribution and for the signal+background distribution taking into account the uncertainty in the scaled expected background. The expected significance is now calculated to be 2.12σ which is slightly higher then the 2.04σ we would get if we used the unscaled background estimate with a 10% relative uncertainty.

C. Profile likelihood test-statistic

Now, we can use the likelihood ratio as the test-statistic as described in section III C. The likelihood functions for the two hypotheses becomes (using $\alpha = 1$):

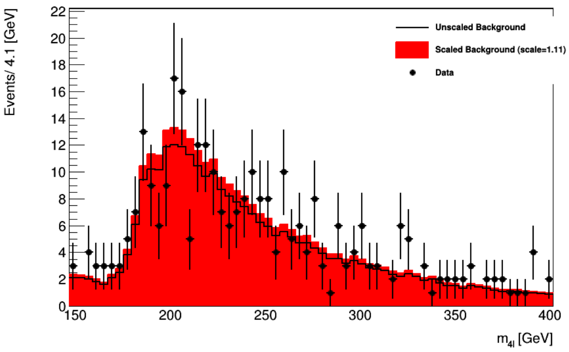


FIG. 5: The rescaled background in the side-band region using the fitted parameter $\hat{\alpha} = 1.11$. The black line shows the unscaled expected background.

TABLE II: The expected significance calculated using 10^6 toys. The unscaled significance is calculated with a relative uncertainty of 10%.

	p-value	significance
Unscaled ($\alpha = 1.00$)	0.020	2.04σ
Scaled ($\alpha = 1.11$)	0.017	2.12σ

$$-2\ln L(\mu = 0) = -2\ln\sum_i f(n_i|b_i) \tag{17}$$

$$-2\ln L(\mu = 1) = -2\ln\sum_i f(n_i|s_i + b_i) \tag{18}$$

where f is the Poisson distribution, n_i is the observed number of events in the data, s_i and b_i are the expected number of signal and background events. The observed test-statistic for the real data is found to be,

$$t = -11.532 \tag{19}$$

We can now find the test-statistic distribution for both hypotheses:

1 A pseudo dataset is generated by drawing a random Poisson number in each bin, using the bin content as the central value.

2 With the background scale factor fixed, $\alpha = 1$, the likelihood functions for the null hypothesis($\mu = 0$) and the alternative hypothesis ($\mu = 1$) is calculated using equations 17 and 18.

3 The test statistic for is calculated using equation 10 and this value is stored in an histogram.

Figure 6 shows the two distributions of the likelihood ratio test-statistic. The distribution for the alternative hypothesis is shown in red, and the black is the distribution for the null hypothesis, also shown is the 1σ and 2σ

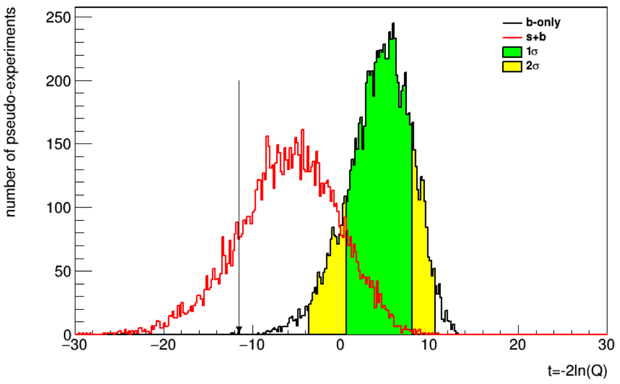


FIG. 6: The distributions of the test-statistics for the two hypothesis, calculated using 10^4 toys. Also shown are the 1σ(green) and 2σ(yellow) bands for the background hypothesis. The value of t_{obs} found from data is marked with an arrow.

TABLE III: The quatiles for the two test-statistic pdf's.

Hypothesis	2σ	1σ	median	-2σ	-1σ
H_0	-3.58	0.76	4.65	7.93	10.56
H_1	-18.11	-11.52	-5.70	-0.35	4.37

fluctuations for H_0. The median values for these distributions are listed in table III. The observed test-statistic is shown as an arrow in the figure.

D. Discovery and exclusion

The confidence levels for the median b-only experiment, the median signal+background experiment and for the data can be found in table IV.

The expected confidence for the median signal+background experiment under the null hypothesis $(1-CL_b)$ is approximately 0.0064(2.49σ). This indicates that while there is a small excess, we cannot expect to reject the null hypothesis and so no discovery is expected. The observed confidence for the data is slightly higher at $p = 0.0002(3.54\sigma)$, yet this is not close enough to the 5σ needed to reject the null hypothesis. So no discovery is made.

TABLE IV: The confidence levels. Discovery: reject the null hypothesis $1 - CL_b < 2.87 \times 10^{-7}$ and exclusion: reject the alternative hypothesis $CL_{s+b} < 0.05$

Confidence level	Median b-only	Median s+b	Median Data
1-CL$_b$	0.49490	0.00640	0.000200
CL$_{sb}$	0.01921	0.50765	0.84425
CL$_s$	0.03802	0.51092	0.84442

TABLE V: The confidence levels for expected discovery/exclusion calculated for 3 different luminosity scale factors.

Scale	$1 - CL_b$	CL_{sb}
1.0	2.49	2.07
1.5	2.95	2.46
2.0	3.29	2.80

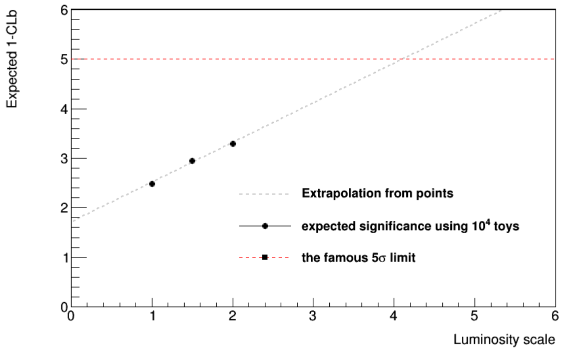


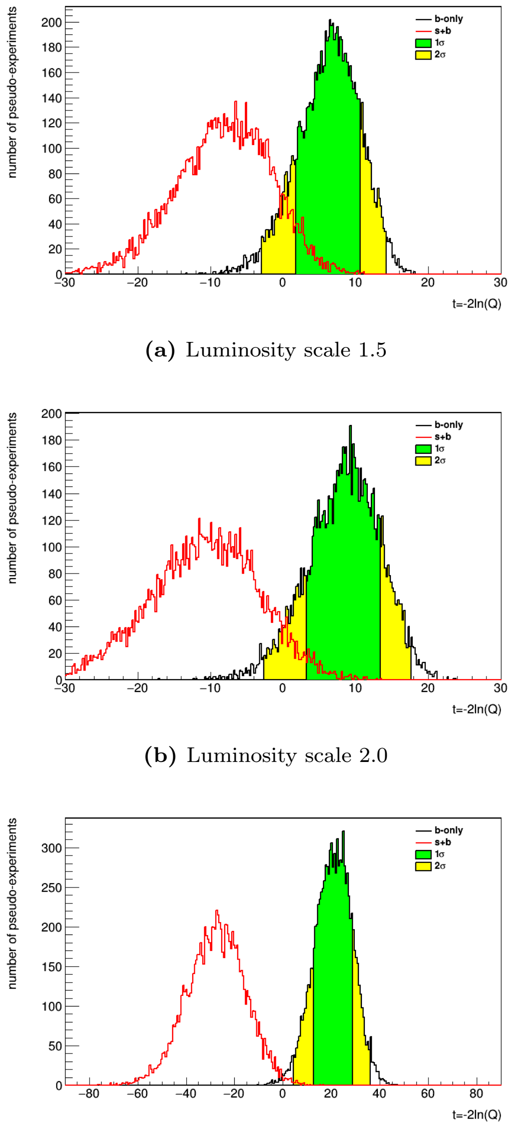
FIG. 7: Extrapolation of the expected significance for discovery. The gray line is the fit and the red line shows the 5 sigma level.

The expected confidence for the median b-only experiment under the alternative hypothesis (CL_{sb}) is 0.0192 which is smaller than the required $CL_{sb} < 0.05$. This means that using this PCL method we are expected to be able to reject the signal+background hypothesis for this mass. But looking at the data, the observed confidence is 0.84, which is too large for an exclusion. In the data $CL_s \approx 0.84 \not< 0.05$ so we cannot exclude a signal at 95% confidence level, as we would have expected with the confidence ratio for the background only experiments being $CL_s = 0.038 \not< 0.05$.

By scaling to 1.5 and 2.0 times the luminosity, we can try to extrapolate to find the luminosity we need to expect to make a discovery. Table V shows the values of $1 - CL_b$ and CL_{sb} used in the interpolation. Figure 7 shows these points and the extrapolated line used to estimate what luminosity we need to reject the null hypothesis. We can expect to make a discovery at about 4.11 times the luminosity. The test-statistic distributions for the scaled cases are shown in figure 8, and we can see that with higher luminosity the two distributions separate. The figure also shows the distributions for 5 times higher luminosity where we can clearly see the separation.

E. Measurement of the production cross section

By repeating the likelihood fit method used for the sideband fit, but scanning the whole mass spectrum, we can find the optimal estimates for both signal and back-



(a) Luminosity scale 1.5

(b) Luminosity scale 2.0

(c) Luminosity scale 5.0

FIG. 8: The test-statistic distributions using 1.5, 2.0 and 5.0 times the luminosity(a,b,c). The distributions are calculated using a rebin factor of 10, with 10^4 toys.

ground scale-factors. We leave both both α and μ as free parameters in equation 4, and calculate $-\ln L(\alpha,\mu)$. The contours of this plot are shown in figure9. The minimum of the log-likelihood is shown as a dot in the center and corresponds to the best fitted values of the parameters,

$$\hat{\mu} = 1.285^{+0.18}_{-0.15} \quad \text{and} \quad \hat{\alpha} = 1.105^{+0.03}_{-0.03} \tag{20}$$

Figure 10 shows the projections on the signal scale factor and background scale factor respectively. These are used to calculate the 1σ errors on the fitted parameters.

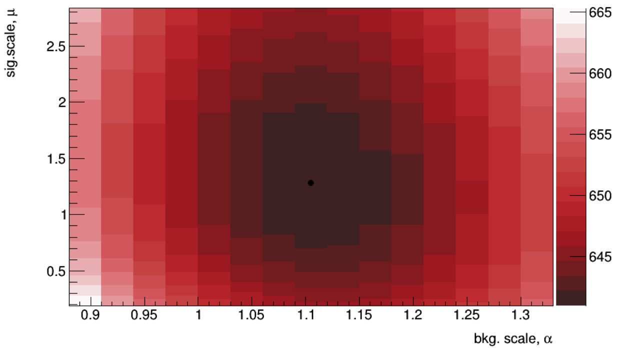
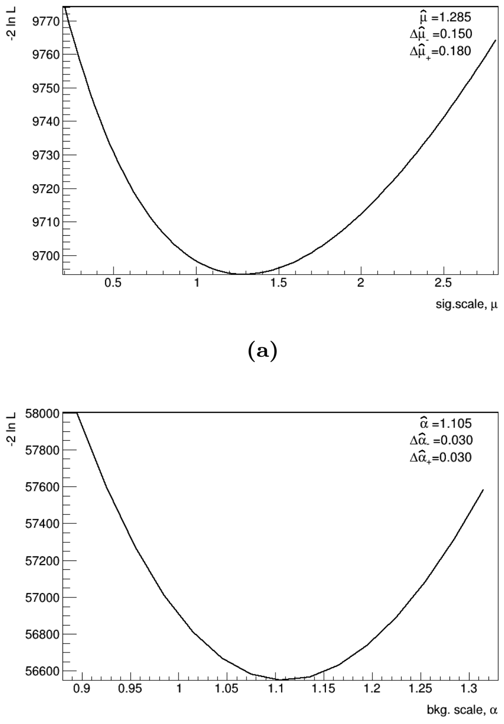


FIG. 9: The contours of $-2\ln L(\alpha,\mu)$. The background scale factor α is on the x-axis and the signal scale factor μ is on the y-axis. The minimum corresponding to $-2\ln L(\hat{\alpha},\hat{\mu})$ is shown with a black dot.



(a)

(b)

FIG. 10: The projections of the log-likelihood function on a) the signal scale factor and b) the background scale factor.

IV. CONCLUSION

The background scale factor α found both from the sideband fit is close to 1, which shows an alright agreement with the data, indicating that the shape of the MC simulations can be trusted. Even though we can see signs of an excess in the mass range $121-129$, we were not able to make a 5σ discovery or exclusion of a $m_H = 125$ GeV Higgs with the current luminosity.

[1] L. Lista, *Statistical methods for data analysis in particle physics; 2nd ed.*, Lecture notes in physics (Springer, Cham, 2017), URL https://cds.cern.ch/record/2293457.

[2] O. Behnke, K. Kröninger, G. Schott, and T. Schörner-Sadenius, *Data analysis in high energy physics: a practical guide to statistical methods* (Wiley-VCH, Weinheim, 2013), URL https://cds.cern.ch/record/1517556.

[3] A. L. Read, *Modified frequentist analysis of search results (the CL_s method)*, CERN-OPEN-2000-205 (2000), URL https://cds.cern.ch/record/451614.

[4] E. Gross, *LHC Statistics for Pedestrians* (2008), URL https://cds.cern.ch/record/1099994.

[5] E. Gross, *Statistics for High Energy Physics. Statistics for High Energy Physics* (2018), URL https://cds.cern.ch/record/2315604.