Optimization of the statistical analysis in a search for Contact Interaction in the dimuon final state

Kamiel Janssens

Supervisors: Jan-Frederik Schulte and Santiago Folgueras

August 30, 2017

Abstract

This project studies search for Contact interaction, a theory beyond the standard model which predicts that quarks are no elementary particles. The constituents are bound together by a new strong force with a characteristic energy scale Λ. An optimisation of the statistical analysis is performed for contact interaction in the dimuon final state. The purpose is to optimise the analysis by comparing the discriminant power of different variables and the performance of the techniques used in the analysis. As a figure of merit the significances of different hypotheses were calculated using a profile likelihood technique. Both a shape-analysis as a single-bin analysis were performed which showed better performance from the shape analysis. Depending on the exact contact interaction model and the characteristic energy scale of the contact interaction either the invariant mass or the angular distribution yields the highest significance.

1 Introduction

This summer student project consists of the optimisation of the statistical analysis for Contact Interaction in the dimuon final state. This study is relevant because we are trying to look at a theory beyond the standard model. Either the theory can be confirmed be the data or it can be excluded to a certain extend. The goal of this project is to optimize the analysis rather than looking to data and testing the theoretical models. Different methods and variables will be used to determine which is the most discriminating way to perform the analysis, which can be applied on similar analyses in the future.

1.1 Contact Interaction

Contact Interaction (CI) is a theory which predicts that quarks are no longer elementary particles but that it are composite particles[1, 2]. The constituents are often called 'preons'. The preons in the quarks will interact with each other by a new strong force. This force is only visible at energies larger than the characteristic energy-scale Λ. For lower energies the preons are confined. However the interactions cannot be observed directly for low energies, they can be observed as a four fermion interaction as shown in Figure 1. This is very much comparable to the Fermi-model for the weak interaction below the mass of the W- and Z-boson.

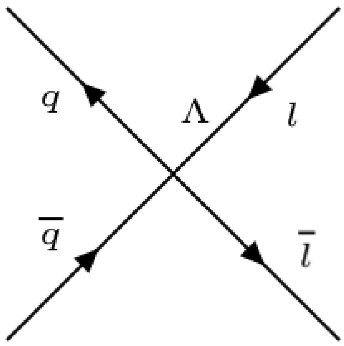


Figure 1: The feynman diagram of the four fermion contact interaction as seen below energy Λ

The Lagrangian of this process is given by Equation 1[1, 2]. There is a term with only left-handed particles, one with only right-handed particles and a third term which contains both left- and right-handed particles. This term has a factor two since it contains both the possible combinations Left-Right and Right-Left. The constants η_{ij} have to be ± 1 or 0. The subscripts L and R refer the respectively the Left- or Right-handedness of the fermions.

$$\mathcal{L} = \frac{g^2}{2\Lambda^2} \cdot (\eta_{LL}\bar{f}_L\gamma^\mu f_L\bar{f}_L\gamma_\mu f_L + \eta_{RR}\bar{f}_R\gamma^\mu f_R\bar{f}_R\gamma_\mu f_R + 2\eta_{LR}\bar{f}_L\gamma^\mu f_L\bar{f}_R\gamma_\mu f_R) \quad (1)$$

In this project we are studying CI with a dimuon final state. This final state is indistinguishable of the Drell Yan (DY)process: $Z \rightarrow \mu\mu$, however one can measure the interference between both when calculating the cross sections as illustrated in Equation 2 by the differential cross section.

$$\frac{d\sigma^{CI/DY}}{dM_{\mu\mu}} = \frac{d\sigma^{DY}}{dM_{\mu\mu}} - \eta_{ij}\frac{I}{\Lambda^2} + \eta_{ij}^2\frac{C}{\Lambda^4} \quad (2)$$

In Equation 2, $M_{\mu\mu}$ is the invariant mass of the dimuon system, I is a term due to the interaction between the standard model DY process and the CI process. The last term in C only depends on CI. Therefore can be concluded that $\eta_{ij} = +1$ corresponds to destructive interference between CI and DY, while a $\eta_{ij} = -1$ corresponds to constructive interference.
Apart from the interference there are other different models of CI that can be studied depending on the chirality.

- LL: $\eta_{LL} = \pm 1$, $\eta_{LR} = \eta_{RR} = 0$
- LR: $\eta_{LR} = \pm 1$, $\eta_{LL} = \eta_{RR} = 0$
- RR: $\eta_{RR} = \pm 1$, $\eta_{LL} = \eta_{LR} = 0$

We define the names for the six categories as follows: ConLL, ConLR, ConRR, DesLL, DesLR and DesRR. The first three have constructive interference the last three destructive. The LL,LR and RR indicate the handedness. Of course combinations are also possible and they will give other models, but they are out of the scope of this project.
In this project models were used with different characteristic energy-scale Λ: 10 TeV, 16 TeV, 22 TeV, 28 TeV and 34 TeV. The main Standard Model background of CI in the dimuon final state is DY. All other backgrounds were neglected during this analysis.

Previous results set limits on contact interaction with both fermions being left-handed using data with a centre of mass energy $\sqrt{s} = 7$ TeV. At a confidence level of 95% the model was excluded for respectively destructive and constructive interference up to 9.9 TeV and 14.3 TeV.[3] This means that with the current centre of mass energy of $\sqrt{s} = 13$ TeV and a higher integrated luminosity the limits now are most likely to be even stricter.

2 Analysis

2.1 Variables

In this analysis several variables to discriminate between signal and background were studied.

- $\cos(\theta^*)$: θ^* is the scattering angle in the dimuon rest frame.
- χ: $\exp(\mid y_1 - y_2 \mid)$, y_1 and y_2 rapidity leading muons ($\chi = (1+ \mid cos(\theta^*) \mid)/(1- \mid cos(\theta^*) \mid)$
- $m_{\mu\mu}$: invariant mass of two leading muons

The invariant mass will be further categorised in four different divisions. The first is the inclusive invariant mass which includes all the events. Afterwards a subdivision is made on the pseudo rapidity η of the muons. A muon with $\eta < 1.2$ is detected in the barrel of the detector, otherwise in the endcap. We define a category bb where both muons are in the barrel of the detector and be which contains all the other cases. As third subdivision $\cos(\theta^*)$ is used. We define two subcategories for the invariant mass CSPos and CSNeg which contain respectively only events with positive or negative $\cos(\theta^*)$.

2.2 Samples

This analysis is fully based on Monte-Carlo (MC) simulated samples. Only a Drell-Yan (DY) background is used in this analysis which is the largest background. The 'Background-like' DY is generated using Powheg upto Next-To-Leading-order (NLO) Monte-Carlo simulation. These sample is afterwards normalised with NNLO cross sections.
Since the signal of CI interferes with the DY background the signal sample need to be generated together with DY. Therefore the signal sample is DY+CI. To avoid many DY events around the Z-peak resonance (91,19 GeV) a mass cut is applied to this sample: $m_{\mu\mu} > 300$ GeV. Otherwise there would be to many events at the Z-peak-value in comparison to the high mass tail of the distribution. The sample is generated using Pythia upto LO Monte Carlo. To ensure we can use this LO order samples and compare them with the NNLO background sample we apply a scaling factor: k-factor = 1,3. Since DY is included in are signal sample also a 'signal-like' DY sample has to be generated to be able to use in the analysis. This 'signal-like' DY sample is generated in exactly the same way as the signal sample except there is no signal and only DY in this sample.
The mass cut which was necessary for the signal sample was afterwards applied on all the samples.

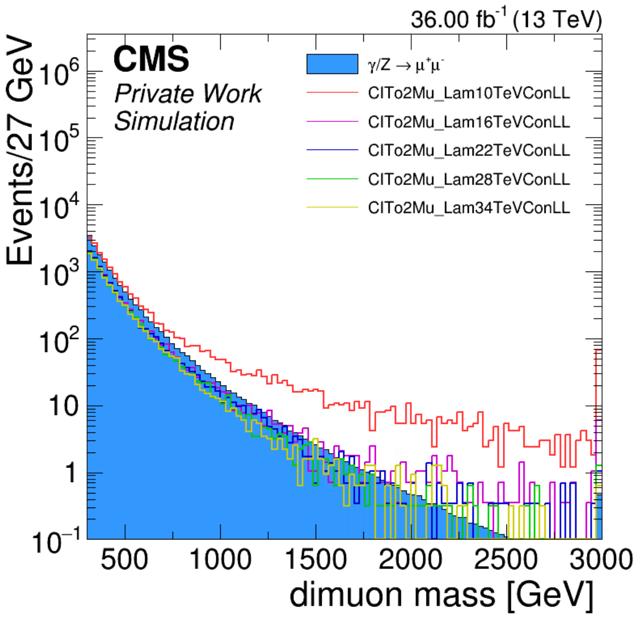
In Table 1 the total number of events of $m_{\mu\mu}$ after passing the mass cut are displayed . There is a clear mismatch between the 'Background-like' DY sample and the 'signal-like' DY sample. However after the correction using the k-factor of 1.3 the values are comparable. The remaining inconsistency is approximately 2.5%.

	'Background-like' DY	'signal-like' DY	'signal-like' DY $\cdot$ k
Inclusive	14609	10966	14256

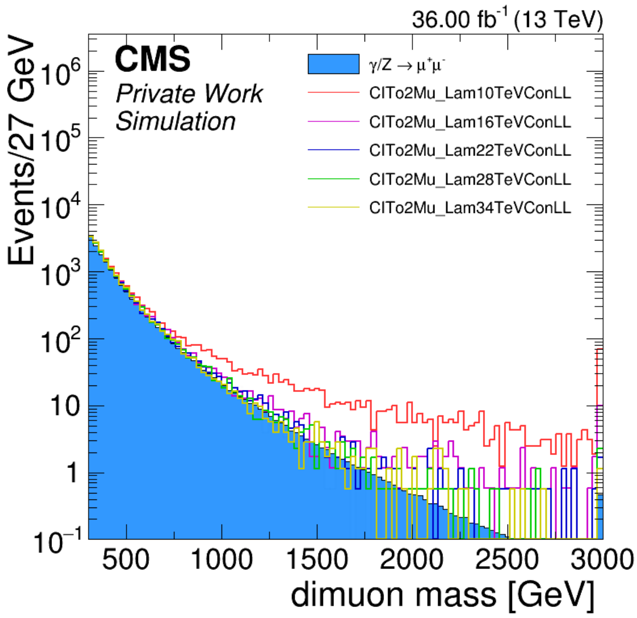
Table 1: Number of events $m_{\mu\mu}$ ($m_{\mu\mu} > 300$ GeV)

We expect that for low invariant mass the signal is negligible and there is only DY. This is because the CI only occurs at an energy-scale Λ at high energies, while DY dominates the low invariant mass regions, closer to the resonance of the Z-boson. This means that the signal-sample only contains DY at low energies and should match perfectly the DY-sample. However this seems not to be the case for the signal samples as seen in 2a. For the 10 TeV sample there is a decent match between the single sample and DY but not perfect. For CI-hypothesises with larger Λ there is a clear inconsistency. Here we have to notice that the cross sections for the 28 TeV and 34 TeV samples were not yet provided, so these are expected to be wrong. They were give values similar to the cross sections of the 16 TeV and 22 TeV sample.
The mismatch between the background can also be seen in the ratio-plots 2c and 2d for the different signal-hypotheses. For $\Lambda = 10$ TeV the ratio is decent, for the other samples the signal is systematic lower than the background.
To correct for this we apply a scaling factor on all the cross sections as seen in 2b. To calculate the cross sections we use a region of low invariant mass were we assume there is no signal but only DY present. We used the region 300 GeV $< m_{\mu\mu} <$ 350 GeV. For all the signal samples and the signal-like DY sample rescaling factors were calculate to match their number of events in this region with the number of events of the background-like DY sample. If we look at 2a and 2b the assumption seems to be correct for samples with $\Lambda \geq 16$ TeV, however the 10 TeV sample has a longer tail of signal towards low invariant mass so it is possible their is signal present in the region used to rescale. Even if the rescaling was done in a region with lower invariant mass (300 GeV $< m_{\mu\mu} <$ 325 GeV) a similar result was found. Using a region with

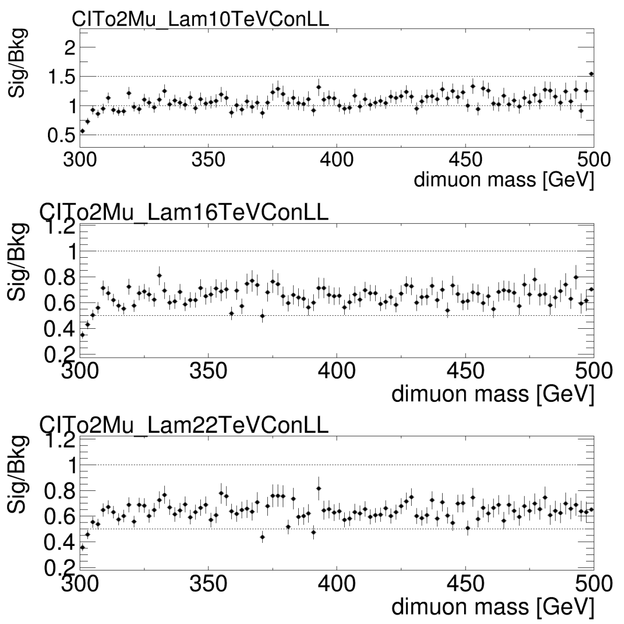
lower invariant mass is not possible since the samples start form $m_{\mu\mu}$ = 300 GeV. Despite the fact that the rescaling probably was not justified for Λ =10 TeV, from a physics standpoint it is of less important. The signal with constructive interference and both fermions left-handed was already excluded up to 14 TeV[3] Therefore Λ =10 TeV is not a physical realistic hypothesis and it is of less important that the rescaling maybe is not completely correct.



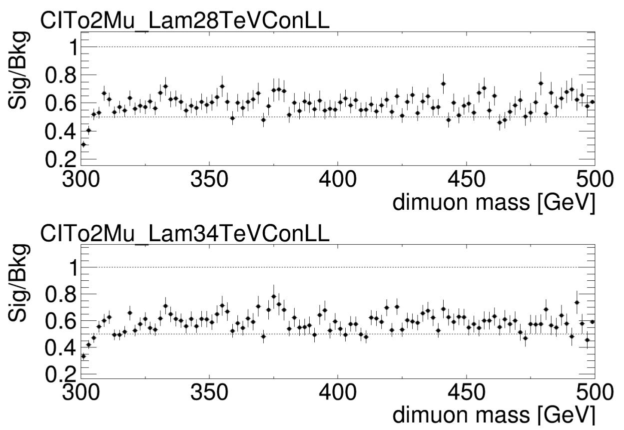
(a) Using unscaled cross sections



(b) Using rescaled cross sections



(c) Using unscaled cross sections



(d) Using unscaled cross sections

Figure 2: The histogram of the invariant mass $m_{\mu\mu}$ (a) before and (b) after rescaling the cross sections. (c) and (d): the ratio of signal over background for the unscaled cross sections for low $m_{\mu\mu}$. For low invariant mass there a mismatch between signal andthe DY background can be seen in (a). This mismatch is also visible in the ratios in (c) and (d).

2.3 Methodology

Afterwards the significance of the signal was studied. The significance was computed using a Profile-likelihood method [4]. The idea of this statistical method is as follows:
You start from a set of parameters 'θ' and you want to estimate these parameters based on measurements 'x'. These measurements follow a probability model $P_x(x|\theta)$. Since the model never is perfect there are systematic uncertainties introduced on the values of the parameters. To reduce these systematic uncertainties nuisance parameters (ν) are added. These additional parameters are related and not independent of the so called parameters of interest θ. However this may decrease the systematic uncertainty there will be larger statistical uncertainties. Instead of using all the nuisance parameters as well as the parameters of interest the equivalent profile-likelihood technique uses only θ where the nuisance parameters are 'profiled'. This means that they are chosen so that the likelihood becomes maximal for the given θ.
To perform this calculations the Higgs combination tool is used [5, 6].

To calculate the significance of the processes we assume the signal is present and the 'measured' data in this case is just the sum of the background and the signal. Since all other background processes are neglected, the only background is Drell-Yan. To calculate the number of signal events the signal is extracted from the signal samples, since there is Drell-Yan present in this sample. For this the 'Signal-like' Drell-Yan sample is used. Afterwards this signal is compared to the 'Background-like' Drell-Yan background to calculate the significance.
The significance was calculated in two difference ways. First a shape-analysis was used and afterwards the results were compared to a single-bin analysis. For a single-bin analysis only one bin is used and the significance is determined according to the total amount of events of background and signal in comparison to the data. For a shape-analysis the studied variable is divided in a predefined set of bins and the Higgs combination tool takes this shapes also in account. As signal and background for instance have a difference shape the significance will increase because the fit is able to use this information. Therefore one expect that the shape-analysis gives better results than a single-bin analysis, which was studied in this analysis.
For the shape analysis the binning of the variables was chosen as follows:

- $\cos(\theta^*)$: 10 equally spaced bins: -1:0.2:1
- χ : 30 equally spaced bins: 0:1:30
- $m_{\mu\mu}$[GeV] : [300,500,700,1100,1900,3500]

The significance was determined on samples using different cuts on the invariant mass. Only a cut on the mass was applied and the number of events for the angular variables was recalculated using this cut. If the number of signal events was negative (there were more events in the signal-like DY sample then in the signal sample) it was set to zero to calculate the significance. As can be seen in Figure 2, the signal to background ratio is higher for high invariant masses. Therefore different mass cuts are applied to investigate if this improves the significance or not. In Table 2 both the cuts and the total number of 'Background-like' DY events (rounded to closest integer) are displayed to give an idea of the amount of statistics for a given cut on the invariant mass.

$m_{\mu\mu} >$ [GeV]	#events 'Background-like' DY
300	14609
500	2684
1000	189
1100	127
1200	85
1300	58
1400	41
1500	29

Table 2: Number of 'Background-like' DY events $m_{\mu\mu}$ (rounded to closest integer) for the cut-flow on the invariant mass $m_{\mu\mu}$

The significance is then evaluated including both statistical and systematic uncertainties. To include all of the systematic uncertainties would lead this project to far. Therefore we worked gradually. First the statistical uncertainties are included. This lead to an uncertainty given by: $1 + \frac{\sqrt{\#events}}{\#events}$. The statistical uncertainty was applied on both signal and background. Secondly two systematic uncertainties were included and as one should expect the significances decreases when adding more uncertainties. For the first systematic uncertainty the invariant mass is scaled by $\pm 1\%$. The second systematic uncertainty includes an uncertainty on the resolution of the invariant mass. This is done by a 20 % Gaussian smearing. The mass resolution differs for barrel or endcap muons, due to the lower granularity of the endcap of the CMS detector.

3 Results

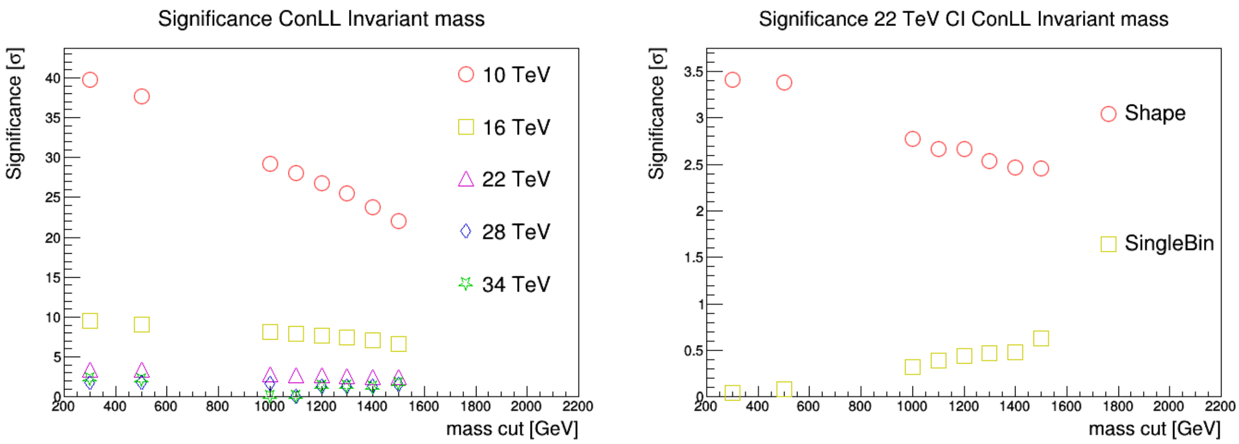
Four different comparisons will be made. First of all the comparison between the different energy-scales Λ of the signal-hypothesis is studied. Secondly a shape-analysis and single-bin analysis are compared. Thirdly a comparison is made between the studied variables for all the different energy-scales for a signal-hypothesis with constructive interference and both fermions left-handed. Finally the different sample-hypothesis are studied for a given energy-scale to determine if there is a difference depending on the destructive or constructiveness as well as the handedness of the signal.

3.1 Different energy-scale

In Figure 3a the significance of the inclusive invariant mass is plotted for the signal ConLL for the different energy-scales Λ. The ConLL sample is the sample where the contact interaction is constructive and both fermions are left-handed. In Figure 2b the signals with lower Λ were more distinguishable from the DY background. This finding is confirmed by the significances. For lower Λ the significance is higher and therefore the signal is more likely to be detected if it was present.
As also could be concluded from figure 2b, there is no large difference between the calculated significances for $\Lambda = 22$ TeV, 28 TeV or 34 TeV. This means that it can be difficult to distinguish between the different hypotheses if the data should follow one of the hypotheses.
Furthermore we notice that the significance of the invariant mass only decreases if the cut on the invariant mass is increased.

3.2 Shape-analysis versus single-bin analysis

In Figure 3b the significance of the inclusive invariant mass from a signal sample ConLL with $\Lambda = 22$ TeV is given. this figure makes the comparison between a shape-analysis and a single-bin analysis. Since the single-bin analysis uses only a single bin these values are independent of the variable used to calculate the values as long as the variable contains the complete dataset, which is not the case for the divided samples of the invariant mass.
The shape-analysis is clearly better than the single-bin analysis, as expected. There is a large difference between both and therefore it is advised to always use a shape analysis if possible to improve the calculated significance. Furthermore we notice that the significance increases with increasing cuts on the invariant mass for the single bin analysis. This also could be expected since only the total amount of events is used. In figure 2b, there can be seen that there is more signal relative to background present for higher invariant mass. Therefore a cut on the invariant mass will improve the signal to background ratio and the single-bin significance as well.



(a) (b)

Figure 3: The significance of the invariant mass for the signal sample ConLL: (a) different energy-scales Λ, (b) $\Lambda = 22$ TeV shape-analysis and single-bin analysis

3.3 Different variables

In Figure 4 a comparison is made to test the discriminant power of the considered variables for a signal sample with constructive interference and two left-handed fermions (ConLL). All the different energy-scales are plotted separately.
We mention that sometimes for a given mass cut the calculated significance is zero, however the significance for both a lower and higher mass cut is non zero. This might be a consequence of a lack of statistics in the 'signal-like' DY-sample for high invariant mass.

In Figure 4a we conclude that the most discriminating variable to is the invariant mass for a hypothesis where $\Lambda = 10$ TeV. Both the inclusive invariant mass as the invariant mass split in all four categories as discussed before have very similar values. The combination of all the four categories gives a slightly larger significance.
Furthermore we notice that both the angular distributions have in general an increasing significance for an increasing cut on the invariant mass, but they remain well below the peak value of the invariant mass.

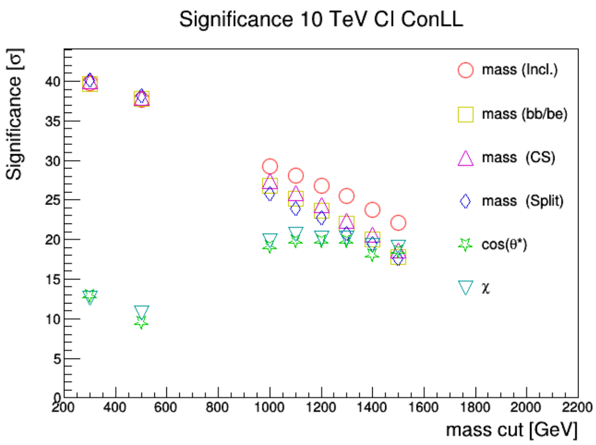
In Figure 4b for a hypothesis of $\Lambda = 16$ TeV very similar conclusions can be drawn. Also here the invariant mass divided in four subcategories yields the best significance. However their is a slight improvement to use a mass cut of 500 GeV instead of the standard mass cut of 300 GeV.

In Figure 4c for a hypothesis of $\Lambda = 22$ TeV some other patterns seem to appear. The subdivided invariant mass is still the most significant variable to use at a mass cut of 300 GeV, but the using the variable χ at a mass cut of 1000 GeV yields a significance which comes close.
Secondly the inclusive invariant mass does no longer yields a very high significance. For lower $\Lambda =$ it was one of the most distinctive variables, but now it yields rather low significances compared to the other variables.

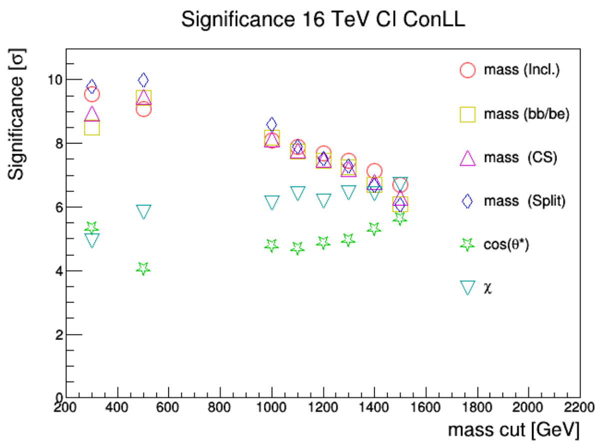
In Figure 4d for a hypothesis of $\Lambda = 28$ TeV the changing pattern becomes more clear. the invariant mass is no longer the most distinctive variable but the angular distribution χ yields the highest significance. The highest yield is for a mass cut of 300 GeV and significance using χ does no longer increase for increasing mass cut.

In Figure 4e for a hypothesis of $\Lambda = 34$ TeV the invariant mass divided in all four categories yields again the highest significance for a mass cut of 300 GeV. The significance using χ with a mass cut of 500 GeV however as a very similar value. The other variables do not yield as favourable significances.

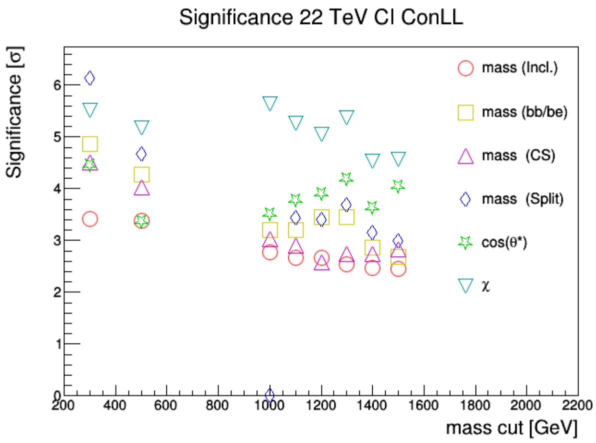
For all the different energies there is at least one variable which yields a significance of 5 σ or higher. This is the discovery limit, but the data matches the background. Based on the limited included uncertainties and neglecting all the backgrounds except for DY, it can be concluded that the ConLL contact interaction up to 34 TeV is not present.



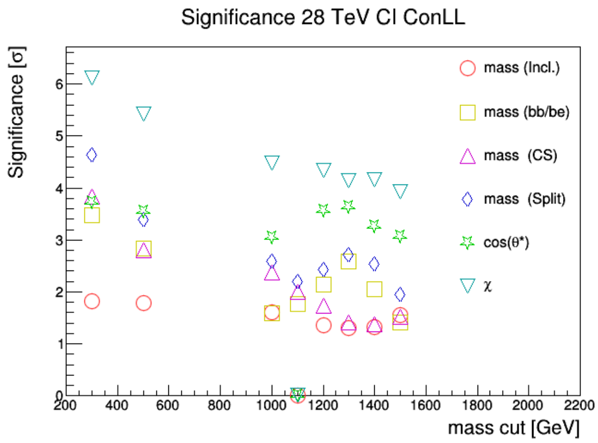
(a)



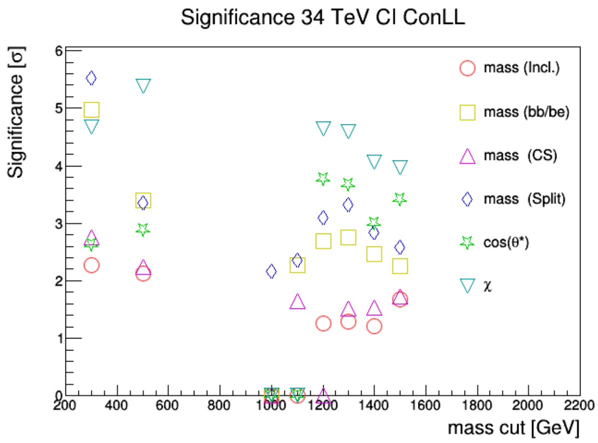
(b)



(c)



(d)



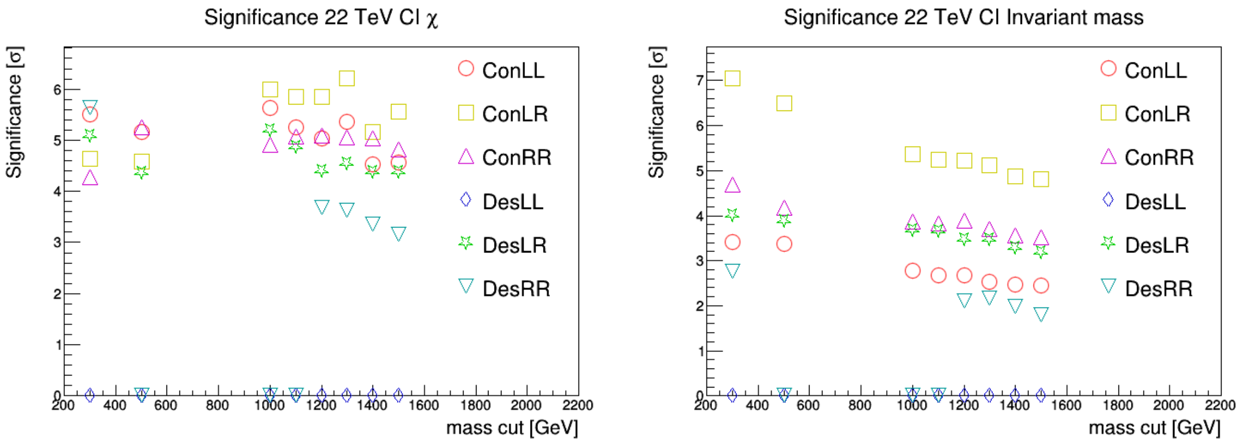
(e)

Figure 4: The significance of studied variables for the signal sample ConLL for $\Lambda =$ (a) 10 TeV, (b) 16 TeV, (c) 22 TeV, (d) 28 TeV and (e) 34 TeV

3.4 Interference and handedness

In Figure 5 the different signal samples are compared for $\Lambda = 22$ TeV and both χ (5a) and the inclusive invariant mass(5b). For the angular distribution χ we can not easily make some general conclusions. The destructive sample with both left-handed fermions has for none of the mass cuts a positive number of events and yields therefore zero significance and can not be detected by looking at an excess of events.

For the invariant mass we the ConLR sample has the highest significance followed by ConRR, DesLR, ConLL, DesRR and the DesLL sample. However it is interesting to notice that for the ConLL sample has a larger significance if χ is used compared to the inclusive invariant mass (not in comparison with the invariant mass divided in the four subcategories). However this is not the case for ConLR and ConRR. For these samples the inclusive invariant mass yields a higher significances then the angular distribution χ.



(a) (b)

Figure 5: The significance of the different signal samples for $\Lambda = 22$ TeV: (a) χ and (b) invariant mass (inclusive)

4 Conclusions

An optimisation of the statistical analysis of Contact Interaction in the dimuon final state was performed. The theory of Contact Interaction predicts that quarks are composite particles. The more fundamental particles (preons) are held together by a new strong force which occurs at energies higher then a characteristic scale Λ. Since the dimuon final state is indistinguishable from the standard model process Drell Yan both processes interfere with each other. In this project different characteristic energy scales were studied as well both destructive and constructive interference between Contact Interaction and Drell Yan.
Different variables were studied to find the variables which yield the best significance and therefore improve our possibility to confirm or excluded a certain hypothesis.

To conclude we summarize the observed findings concerning the cross sections and which techniques and variables are the most efficient to perform the analysis with.
We can conclude the performing a shape-analysis yields higher significances than using a single-bin analysis. They differ by a factor of several units.
Which variable and accordingly which cut on the invariant mass should be used to yield the highest significance is not as clear depending on the situation. For hypotheses with lower Λ the invariant mass divided in four subcategories yields the best results. The categories splits the invariant mass for both muons in the barrel or at least one in the endcaps of the detector, combined with the requirement that the event has either a positive or negative $cos(\theta^*)$. For a hypothesis with a larger Λ the angular distribution χ can yield a comparable or higher significance. Sometimes a larger cut on the invariant mass is required than the standard cut of 300 GeV on all the samples. Finally we mention that these findings also can vary for different signal models, this means other type of interference or handedness of the fermions. Therefore it is advisable to verify which is the best option for the analysis you are performing.

5 Outlook

If in the future someone wishes to continue on this topic it is most of all interesting to redo the analysis using the physical cross sections instead of needing to rescale the samples to ensure the signal samples match the Drell Yan background at low invariant mass.
Furthermore only a limited amount of systematic uncertainties were included. To improve the analysis and make it more physical relevant other systematic uncertainties should be included as well.
Finally other background processes can be included. In this analysis only DY, which is the main background was included. For a more final result also other backgrounds should be included since they will influence the final significance even they are less prominent in this specific final state.

6 Acknowledgements

I would like to thank my supervisors for the help with the project and their efforts to teach me the subtleties of a data-analysis in elementary particle physics. Furthermore I would like to thank CERN to give me the opportunity to be part of the Summer Student Programme.

Bibliography

[1] L. Apanasevich, A. Hinzmann, K. Rabbertz, N. Varelas and J.Zhang, *Search for new physics with dijet angular distributions in proton-proton collisions at* $\sqrt{s} = 13TeV$, CMS Draft Note (2017)

[2] A. Albert, N. Neumeister and J.-F. Schulte, *Interpretations of the Z' → ll search in different physics models*, CMS Draft Analysis Note (2017)

[3] S.Chatrchyan et al (CMS Collaboration) *Search for contact interactions using the inclusive jet* p_T *spectrum in pp collisions at* $\sqrt{s} = 7\ TeV$, Physicaml Review D87, 0572017 (2013)

[4] G. Cowan, *Statistics for Searches at the LHC*(2013) arXiv:1307.2487v1[hep-ex]

[5] N. Wardle, *Documentation of the RooStats -based statistics tools for Higgs PAG*, https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideHiggsAnalysisCombinedLimit, (2017)

[6] F. Ricci-Tam, *Higgs Combination and Properties: Long Exercise - CMSDAS 2014*, https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideCMSDataAnalysisSchool2014 HiggsCombPropertiesExercise, (2014)