Search for $H \to \gamma\gamma$ in CMS Open Data

Christian Staufenbiel
Leibniz Universität Hannover

September 5, 2018

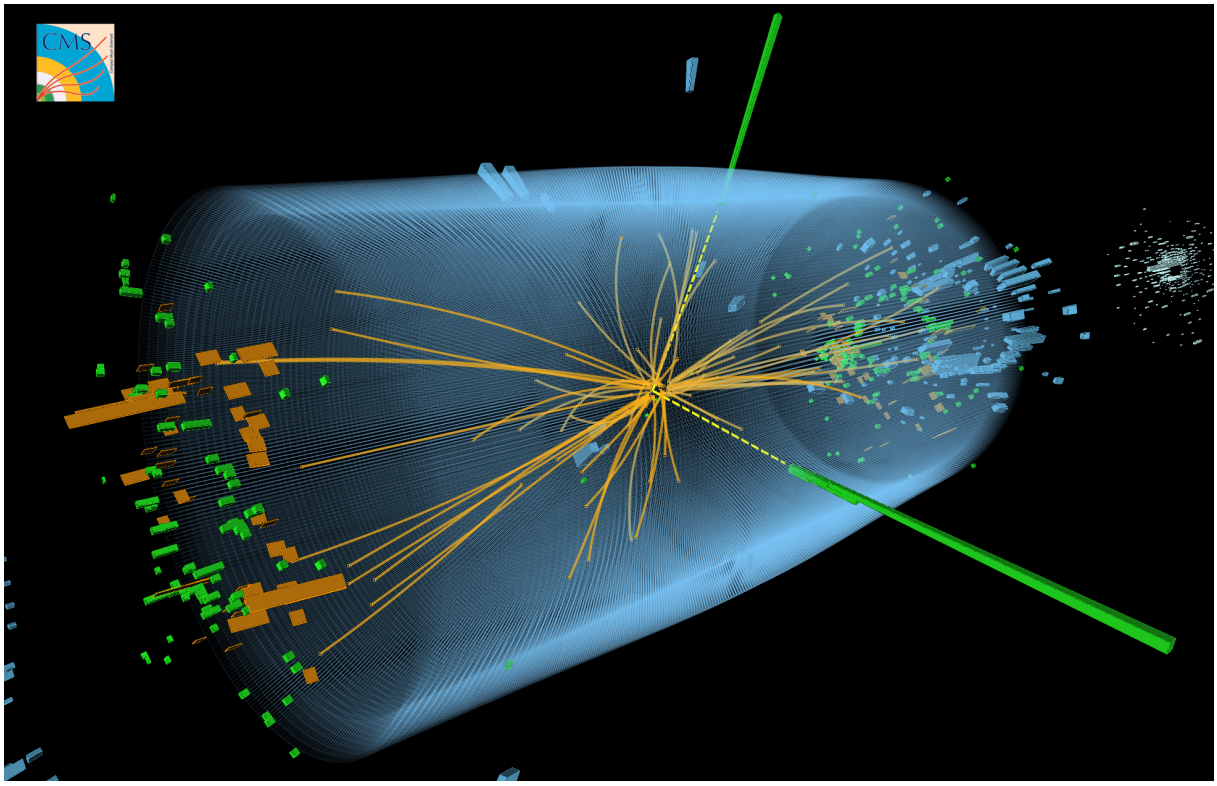


Figure 1: $H \to \gamma\gamma$ event candidate in the CMS detector, which we want to detect in this analysis example. Dashed yellow lines indicate two high energetic photons that can originate from a Higgs boson decay.

Contents

1 Introduction

The analysis of data from high energy particle collisions is a big research area, which need to handle large datasets. Results of analyses on these datasets are constantly published in several journals. In 2014, the observation of the Higgs boson to two photons [8], was one of the biggest results of these analyses.
Over the last years efforts are made to provide these data samples, used in published papers, publicly on the *CERN Open Data portal*. Analysis examples are provided with the datasets to give anyone the ability of performing and developing similar analyses.
In this work we will focus on the analysis of the $H \to \gamma\gamma$ decay channel which was used to find the Higgs boson. A display of a candidate for this event is shown on the title page in fig. 1. We aim to reproduce the results of the published papers [4] [8]. However the time on this summer project is limited and therefore the analysis is simplified in some points.
We provide a code example[1], which is a showcase of accessing and performing an analysis on the datasets.

2 Open Data

The CERN Open data portal[2] provides data, analyses and documentation to the public. It preserves data from experiments and enables public users to perform data analyses on 'real' data from experiments at CERN. The goal is to provide data and analyses for research but also for educational purposes. Therefore data is available from several experiments at CERN. For this specific analysis data from the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider will be used. For the $H \to \gamma\gamma$ channel there are several primary datasets available, which include only events with specific photon triggers fired [1] [2] [3]. These datasets contain $\sim 25TB$ of collision data which are analyzed in this work. For 2012 data a Monte Carlo (MC) simulation is available for the *gluon-gluon fusion* production mode ($\sim 90\%$ of overall production) at $m_H = 125GeV$ [9]. This simulation is used to estimate the predicted signal in the data (without background). However for 2011 there is currently no MC-signal available at the Open Data portal. MC-simulations are used to determine expected results and improve the performance of this analysis.
A *CMS virtual machine*[3] is provided for easy access to the data and all the tools needed for the analysis are pre-installed. These tools include the data analysis framework `ROOT` and the CMS-Software, which is continuously developed in the CMS software repository[4].

3 CMS Detector

A brief introduction is given to the CMS detector as it is crucial to understand the setup of the measuring device when analyzing the collected data. As figure 2 shows, the silicon tracker, electromagnetic (ECAL) and hadron calorimeter (HCAL) are surrounded by a superconducting solenoid, which provides a magnetic field in the inside of the tracker. Hence the tracks of charged particles are bend before they hit the calorimeter. The silicon tracker recognizes hits of particles. From these hits the tracks of particles are reconstructed. The bending of a charged particle track is used to measure the particle momentum. The energy of the particle is deposited in the ECAL or HCAL, which are able to measure the energy deposits. The ECAL consists of 80.000 lead-tungstate crystals, which produce light according to the energy of the hitting particle. The crystals are ordered in grid of squares, where each crystal in the barrel (endcap) has a surface of

[1] https://github.com/cms-opendata-analyses/2011-photon-2012-doublephoton-higgs-hgaga
[2] http://opendata.cern.ch/
[3] http://opendata.cern.ch/docs/cms-virtual-machine-2011
[4] https://github.com/cms-sw/cmssw

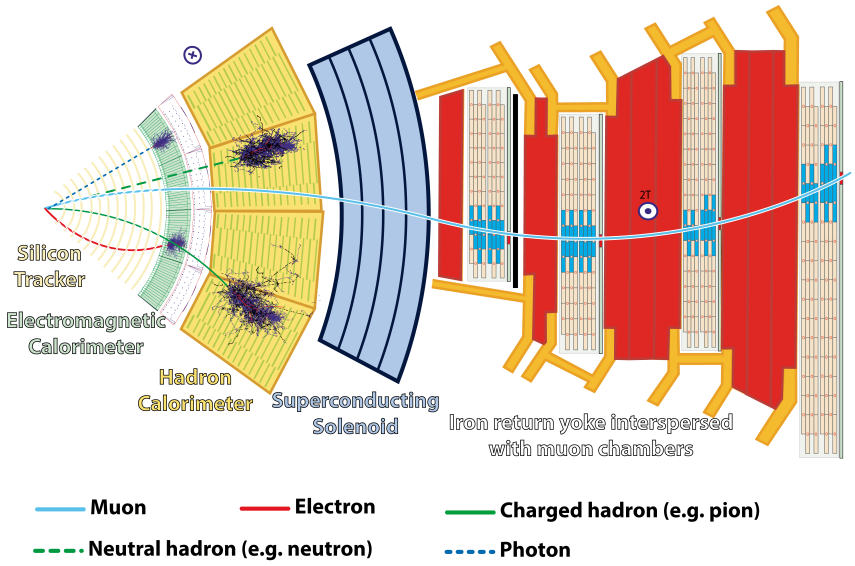


Figure 2: A slice of the CMS detector showing the tracker, calorimeter and muon detector. Taken from: http://cds.cern.ch/record/2120661/

$2cm \times 2cm$ ($3cm \times 3cm$). Therefore the resolution in the barrel is better than in the endcap. The HCAL is made out of brass/scintillators. When hadronic particles interact with these materials, showers are formed. These showers emit light while passing through the different layers of the HCAL. The light of the ECAL/HCAL is used to measure the energy of the incoming particles. For muons the interaction with the calorimeter is too weak and they will travel through the solenoid and can be observed in the muon detector.
The azimuthal position in the detector is described using the pseudorapidity $\eta = -ln[tan(\theta/2)]$, where θ is the angle of a particle trajectory to the beam direction. Thus $|\eta| = \infty$ for particle direction into beam direction (forward) and $|\eta| = 0$ if the particle trajectory is perpendicular to the beam direction.
As the detection devices do not cover the whole range of η with the same setup, the resolution of the data detection changes within the position of the detector. This will effect the further analysis, where a good signal-to-background resolution is required.
A more detailed description of the CMS detector is given in [10].

4 Cut based analysis

The datasets consist of a large number of bunch crossings. In each bunch crossing many pp collisions take place. For every collision several photon candidates are reconstructed. These candidates need to be filtered in such a way, that we keep the photons which could be generated by $H \to \gamma\gamma$ decay. Each photon has specific properties (energy, shower shape, etc.) , which are stored in different variables.
In this analysis we use these variables to decide if a photon candidate should enter the results or if it is discarded. Therefore *cuts* are applied on the variables. This means if the variable does not fulfill the cut criteria its corresponding photon candidate is rejected and can not enter the results.
This cut based selection is used to reduce the background to the $H \to \gamma\gamma$ decay. The reducible background consist mainly of $pp \to \gamma + jet$, $pp \to jet + jet$, where the jets are misidentified as photons. Also there can be multiple π^0 in the jet which decay to two photons. To suppress this background the isolation of the photon from hadronic energy (jets) is used. However there is also an irreducible background from the prompt diphoton production. These photons are

isolated from jets and are mainly rejected by the kinematic selection.
This analysis approach was used for 2011 data [4] and is also described as an alternative analysis to the *multivariate analysis* (MVA) used for 2012 data [8]. Details are described in CMS internal documents [5] [6] [7]
The cut based method is simpler to execute and can give an introduction in handling the data provided by CMS Open Data.
We try to keep our analysis as close as possible to the one performed in the CMS papers. However we have to simplify our analysis in some cases, since the time on this project is limited. The cut values for each class are provided in the Appendix.

4.1 Photon triggers

The datasets used in this analysis contain only events for which one of the photon triggers is fired. This decreases the number of events to be analyzed, since all other kinds of events are not included. We implemented the triggers in the code anyway, such that it can be used when analyzing MC samples and data. We check if there is at least one trigger fired for each event, which is indeed the case. But the implementation can be used to create other trigger selection and demonstrates the use of the trigger system in Open Data.

4.2 Photon classification

Photons have a significant probability to convert to e^-e^+ -pairs before they hit the ECAL, since the tracker has a thickness of about one radiation length of the photon. Unconverted photons are better reconstructed than converted photons. Additionally the reconstruction resolution varies in different areas of the CMS detector, as described before.
Therefore we need to separate photons according to their pseudorapidity η and R_9. The R_9 variable is a crucial to identify the shower shape of a photon when it hits the ECAL. It is defined by $R_9 = E_{3\times3}/E_{SC}$. $E_{3\times3}$ is the energy of the 9 crystals in the ECAL centered around the highest energetic crystal. E_{SC} is the energy of the *super cluster* which is also clustered around the highest energetic crystal (larger than 3×3). If the R_9 variable is higher, the photon tends to be unconverted and therefore its energy is better reconstructed than for photons with low R_9.
The CMS detector is split into two parts according to the pseudorapdity: Barrel and Endcaps. The barrel is located in the range of $|\eta| < 1.4442$ and the endcap in the range of $1.566 < |\eta| < 2.5$. The transition area between barrel and endcap $1.4442 \leq |\eta| \leq 1.566$ is excluded.
With these two variables we define four classes in which each photon is categorized. If a photon is not in the specified range of barrel or endcap it is discarded. As a separation value of the R_9 variable 0.94 is chosen. Hence we create four classes of photons, which have different signal-to-background ratios and are reconstructed with different precisions.

	Barrel	Endcap
$R_9 > 0.94$	Class 1	Class 3
$R_9 \leq 0.94$	Class 2	Class 4

The cut values applied on the variables, which we describe in the following, are optimized for each class to pick the best photon candidates for the $H \to \gamma\gamma$ detection.

4.3 Kinematic selection

As a first selection, we want to filter out low energetic photons of each event, since low energy prompt photons are most likely not $H \to \gamma\gamma$ photons (since we know $100GeV < m_H < 180GeV$). For each event we run through all photons and combine it with all the other photons. $p_{T,1}$ and

$p_{T,2}$ are the transverse momenta of photon 1 and photon 2. The combined energy is calculated from the photons momenta p_1 and p_2.

$$m_{\gamma\gamma} = \sqrt{\eta_{\nu\mu} p^\mu p^\nu} \quad \text{where} \quad p^\mu = p_1 + p_2 \tag{1}$$

Consider $p_{T,1} > p_{T,2}$. We then reject photon 1 if $p_{T,1} < m_{\gamma\gamma}/3$ or photon 2 if $p_{T,2} < m_{\gamma\gamma}/4$. This pair is then rejected, but still each of the photons can pass this selection with another photon. Low energetic photons, which are mainly background, are rejected when the transverse momentum is less than $25GeV$.

4.4 Shower shape

The R_9 variable was already introduced for the photon classification and is also used to reject some photons in the low R_9 classes (Class 2 and 4). The photons R_9 variable should not be below a certain threshold value defined for each class.
The transverse shape of the cluster in the ECAL is defined by the $\sigma_{i\eta i\eta}$ variable.

$$\sigma^2_{i\eta i\eta} = \frac{\sum_i w_i(\eta_i - \bar{\eta}_{5\times5})^2}{\sum_i w_i} \qquad \text{with} \qquad w_i = max\left(0, 4.7 + ln\frac{E_i}{E_{5\times5}}\right)$$

E_i and η_i denote the energy and pseudorapidity of the i^{th} crystal in the 5×5 cluster around the highest energetic cluster respectively. The value of $\sigma_{i\eta i\eta}$ tends to be smaller for photons which are isolated from (hadronic) background, which we expect for photons originating from Higgs decay. Hence we require all photons $\sigma_{i\eta i\eta}$ to be below a certain threshold.

4.5 Tracker and calorimeter isolation

Isolation is a tool to identify non-prompt photons, which originate from 1) γ in jets and 2) jets misidentified as γ. The isolation variables are the sum of transverse energy in a hollow cone deposited in the ECAL or HCAL, namely Iso^{ECAL} and Iso^{HCAL}. The inner most part of the cone is excluded to subtract the energy which originates from the photon itself. The radius of the cone is defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. In the implementation a cone of $\Delta R = 0.3$ or 0.4 is chosen.
For prompt photons, without jet background, we expect the isolation variable to be low due to the narrow shape of the photon hit in the calorimeter. However there are contribution to the sum of E_T in the cone from other collisions in the same bunch crossing. To deal with this issue, we subtract a fraction of the median transverse energy ρ from the isolation sums. The fraction is determined by the effective Area A_{eff}, which is proportional to the calorimeter activated by particles from the same bunch crossing.
The track isolation Iso^{track} is also the sum of E_T in a cone of the ECAL, but only the transverse energy of tracks originating from within a small range around the primary selected vertex contribute to that sum.
The isolation sum is defined as:

$$\sum Iso = Iso^{track} + Iso^{ECAL} + Iso^{HCAL}$$
$$\sum Iso^{PUCorr} = \sum Iso - A_{eff}\rho$$

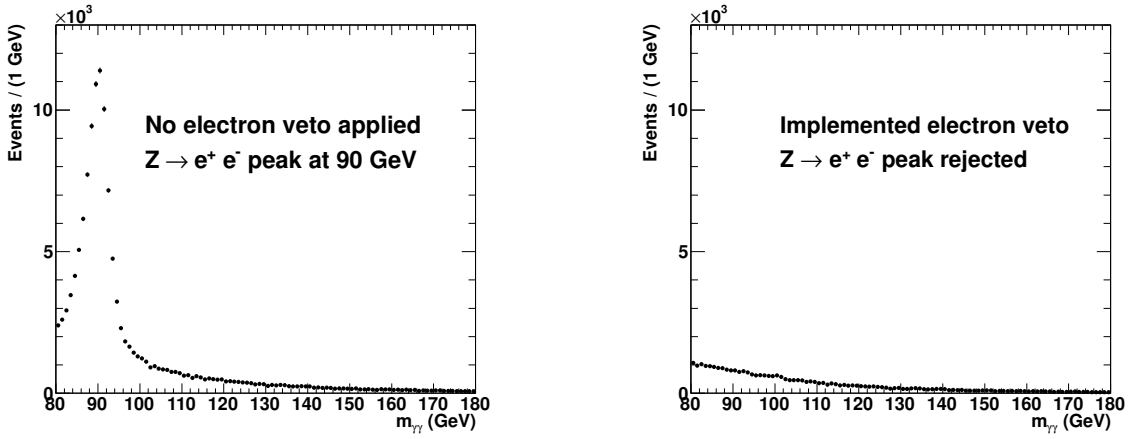
The pile-up corrected isolation sum $\sum Iso^{PUCorr}$ is scaled by the factor $50/p_T$, where p_T is the transverse momentum of the photon. By scaling with this specific factor, the isolation sum increases for lower energetic photons, which then have a higher chance of beeing rejected. For photons with the typical signal photon energy ($\sim 50GeV$), the scaling factor is near to 1.
Additional to the cut applied on the isolation sum, another cut is placed on the track isolation $\sum Iso^{track}$ only, which is also rescaled by $50/p_T$.

4.6 PF Isolation

For 2012 data a new *particle flow* (PF) algorithm was used. Particle flow aims to identify and reconstruct particles from *pp*-collisions using all sub-detectors, which results in a better performance. Therefore the isolation used in 2011 data is replaced by the *PF isolation*. The calorimeter isolation variables still serve as a soft preselection. The particle flow algorithm tags candidates as different particles. Using this tagging, we can define isolation variables for photons and hadrons. Both isolations are (as for calorimeter isolation) the sum of E_T in a specified cone of $\Delta R = 0.4$.
The value of the isolation depends on the selected primary vertex. In 2011 data, the isolation variables are given by the `PhotonCollection` which store all photons for the events. For 2012 data, the particle flow isolation can be calculated for every primary vertex. Therefore the selection of the primary vertex, we pass to the particle flow algorithm, is a crucial part. As the time of this project is limited we choose the first of all primary vertices available for all events. This introduces some errors, which need to be fixed (currently under development). However we can then calculate the PF isolation sums in the cones. Additional to the photon and hadron isolation a cut is placed on the *worst vertex photon isolation*, which refers to the highest photon isolation value, when varying the selected primary vertex. As we have all primary vertices and can calculate the photon isolation for each of these vertices, we can easily implement this cut variable.

4.7 Electron Veto

When applying cuts on the variables above and plot the number of events against $m_{\gamma\gamma}$ (figure 3), we observe a big peak at $90GeV$. This peak arises from the decay of a Z boson to a e^-e^+ pair, which is well known and used to test and calibrate analysis tools for the Higgs analysis in the paper. However these electron/positron-pairs are undesired in this analysis and need to be discarded. As these pairs passed through all the cuts applied beforehand, we need to apply an electron veto, which sorts out all electrons from the possible photon candidates. Therefore we use the super cluster position of the photon candidate and check if there exists an electron candidate in the same super cluster. If this is the case we reject the photon candidate. In the analysis a more sophisticated electron veto is applied, but as shown in fig. 3 our approach is sufficient enough to reject the peak at $90GeV$.



(a) (b)

Figure 3: Effect of the electron veto in 2011 data

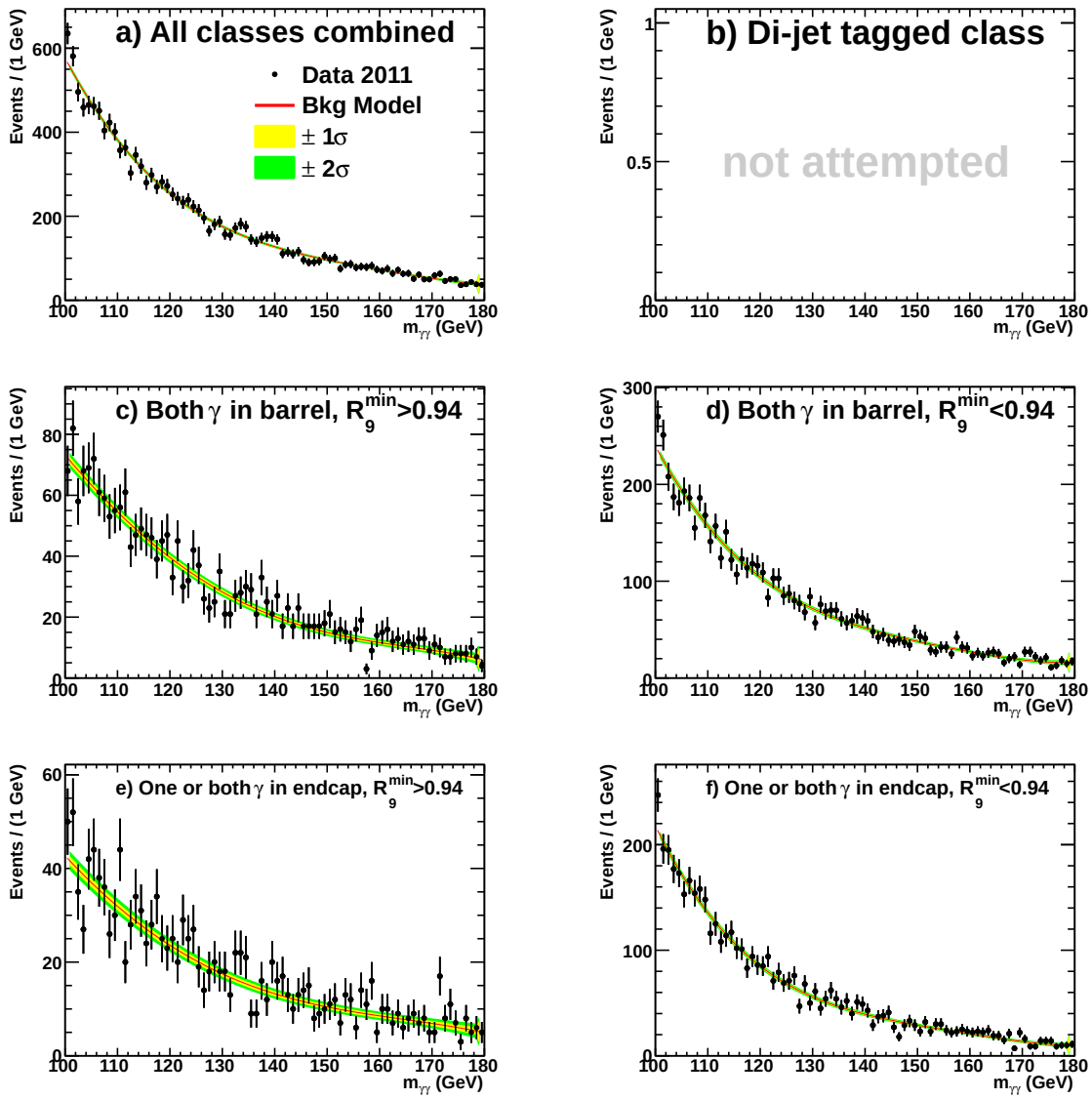


Figure 4: Mass distribution for individual event classes in 2011 data

5 Invariant mass distribution

After applying all the cuts to the dataset we hold a reduced number of possible $H \to \gamma\gamma$ events with a corresponding invariant mass $m_{\gamma\gamma}$. We classify the possible Higgs decay events in four different classes to seperate different events according to their signal-to-background ratio. Tighter cuts are applied in classes with higher signal-to-background ratio than in other classes. For each photon pair we use the smaller R_9 variable of both photons, R_9^{min}, and the position of the photons.

	Both γ in Barrel	One γ in Endcap
$R_9^{min} > 0.94$	EventClass 1	EventClass 3
$R_9^{min} \leq 0.94$	EventClass 2	EventClass 4

The $m_{\gamma\gamma}$ distribution is shown for each class in fig. 4 for 2011 data.
To fit the background we use a 5^{th} order polynomial in the range $100 - 180GeV$. The same function is used in the papers. For signal fitting we add a gaussian peak to the background fit fixed at $\mu = 125GeV$ with width and amplitude as free parameters.

Comparing the plot 4 created from 2011 Open Data to the one published in the 2011 paper [4] (shown in figure 8) shows, that the distribution has the same shape in each class. Since we calculated the integrated luminosity to $2.33fb^{-1}$ (compared to $5.1fb^{-1}$ in the paper) we expect $\sim 50\%$ of events in each class. This is roughly the case for the high R_9 classes. In the classes with low R_9 we see slightly too many events in each bin. This could be a result of the simplified analysis chosen above, which does not follow the paper analysis in every step. Also 2012 data, shown in fig. 7, shows a similar distribution.

6 Significance of 2012 analysis

Since we only have a MC-signal (prediction) for 2012, we calculate the significance of our result only for 2012 data. Therefore a much simpler approach is used than the *profile likelihood ratio* from the paper[8].
The fit $f(x)$ of the data in the range of $100 - 180GeV$ is assumed to describe the background completely, even though a signal could be included somewhere aroud $125GeV$. To calculate the deviation of the data points d_i to the background function $f(x)$ we calculate χ^2 in various ranges around $125GeV$.

$$\chi^2 = \frac{(\sum_i d_i - \int f(x)dx)^2}{\sigma_{int}^2 + \sigma_{data}^2}$$

Here the data points d_i are summed and the background function $f(x)$ is integrated over the specific range chosen around $125GeV$. The integral error σ_{int} is calculated from the error of the parameters given by the fit. For the data error we assume that each bin has an error of $\sqrt{d_i}$. Thus the overall data error squared, σ_{data}^2, is the sum of all data points $\sum_i d_i$.
To add background to the MC signal data points s_i we add the number of events in the background fit at the mass point $d_i = f(x_i) + s_i$. Thus we can compare the results of the MC signal to the real data.
From the χ^2 results we calculate the p-value, to compare it to the paper later on.

$$p = \int_{\chi^2/2}^{\infty} \frac{t^{r/2-1}e^{-t}dt}{\Gamma(r/2)} \qquad \text{with} \qquad r = \text{number degrees of freedom}$$

As we sum all data points d_i into a single bin, we only have one degree of freedom, i.e. $r = 1$.
To retrieve the significance from the p-value we can use a gaussian distribution $g(x)$ with $\mu = 0, \sigma = 1$.

$$1 - p = \int_{-\infty}^{z} g(x)dx \qquad \text{needs to be solved for z.}$$

z gives the factor of how many times the standard deviation σ is exceeded. Figure 5 shows, that we expect (MC signal) the lowest p-value, i.e. highest significance, in the range of $123-127GeV$. The observed p-value (data) is even below that.

$$123 - 127GeV \qquad \text{Expected 'MC'} : 1.23\sigma \qquad\qquad \text{Observed 'data'} : 1.90\sigma$$

However for the observation of the Higgs boson a significance of 5.7σ (expected 5.2σ) was calculated. These results included the complete data set and used the MVA instead of the cut based analysis. Also another method for calculating the p-value was used as stated above.

7 Combined data

To compare the data to published results for the observation of the Higgs boson [8], we combine both datasets into one histogram. As we see in the individual class plots (figure 4 and 7) the

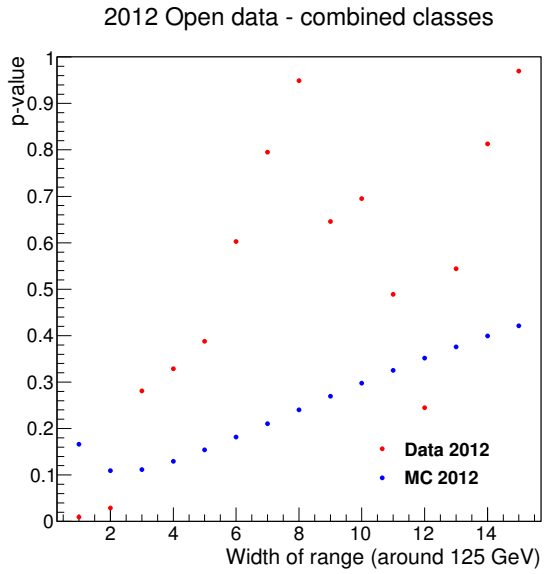
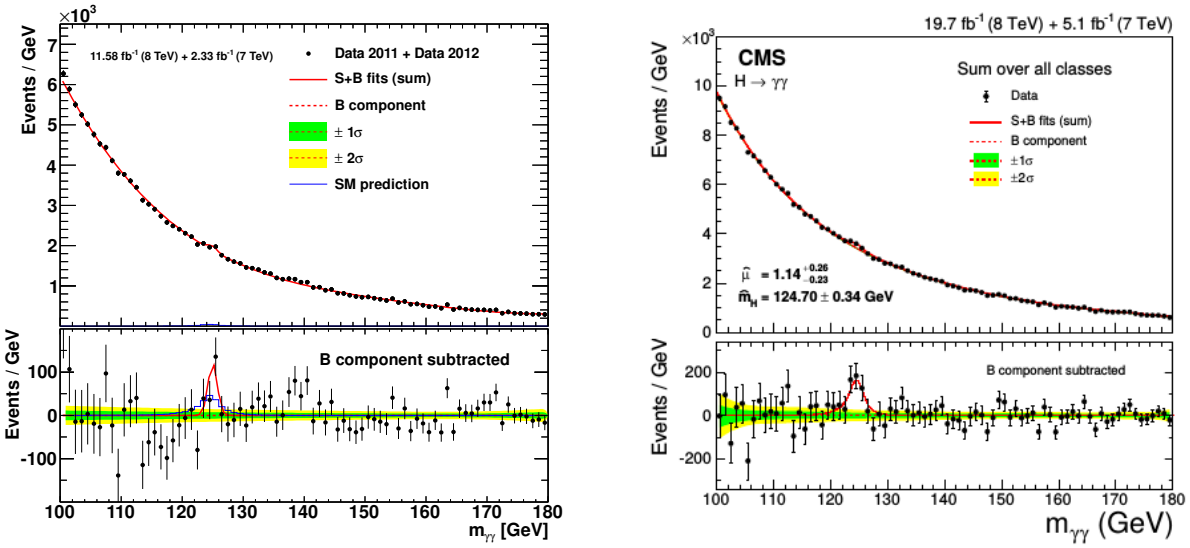


Figure 5: p-value for several ranges symmetric around 125 GeV

2012 dataset includes a larger fraction of events then 2011 data. Thus the combined plot is dominated by 2012 data. Again the peak (at $125GeV$) originates from the additional gaussian



(a) Open Data

(b) Higgs discovery paper [8]

Figure 6: Invariant mass distribution for combined data (2011+2012) with a subtracted background plot.

fit at fixed $\mu = 125GeV$. If we subtract the background function from the data points (lower subpad of figure 6a) a high deviation to the background is observed at $125GeV$. However this single data point is not a sufficient enough to draw any conclusion about the Higgs boson.
Comparing the published result (figure 6b) to our Open Data result both share the general appearance. In the Open Data we calculate the integrated luminosity to $13.9fb^{-1}$ compared to $24.8fb^{-8}$ in the published paper. As we expect in each bin our result holds roughly 50% of the published results. However the peak of the published results is more pronounced.

8 Conclusion

We performed a simplified cut based analysis on the currently available data set and MonteCarlo-simulations in Open Data. Created plots have comparable properties as the published ones. We calculated the significance of our results with a simplified method. However the Higgs boson could not be observed in the $H \to \gamma\gamma$ channel with large significance using the simplified analysis in the Open Data samples.
Improvements can be done (as stated before) in terms of vertex selection in 2012 data and other parts of the analysis. With more data being published in the future, the significance of observing the Higgs boson can increase, without changing the analysis at all.
Additional to the analysis of the Open Data, we provide a code example[5] for accessing and performing an analysis using the Open Data. This will serve researchers and educators as an introduction to the Open Data portal and a first interaction with the analysis of CMS Open Data. A detailed technical description of that code is given in the Appendix.

9 Acknowledgments

I am very happy that I got the opportunity to work in this fascinating and friendly environment at DESY. Special thanks to Oleksandr Zenaiev, who motivated me for this project and helped me a lot with all the problems I encountered working with Open Data. Thanks to the organization team for planning this inspiring and fun time at DESY.

10 Appendix

10.1 Mass distribution plots

[5] https://github.com/cms-opendata-analyses/2011-photon-2012-doublephoton-higgs-hgaga

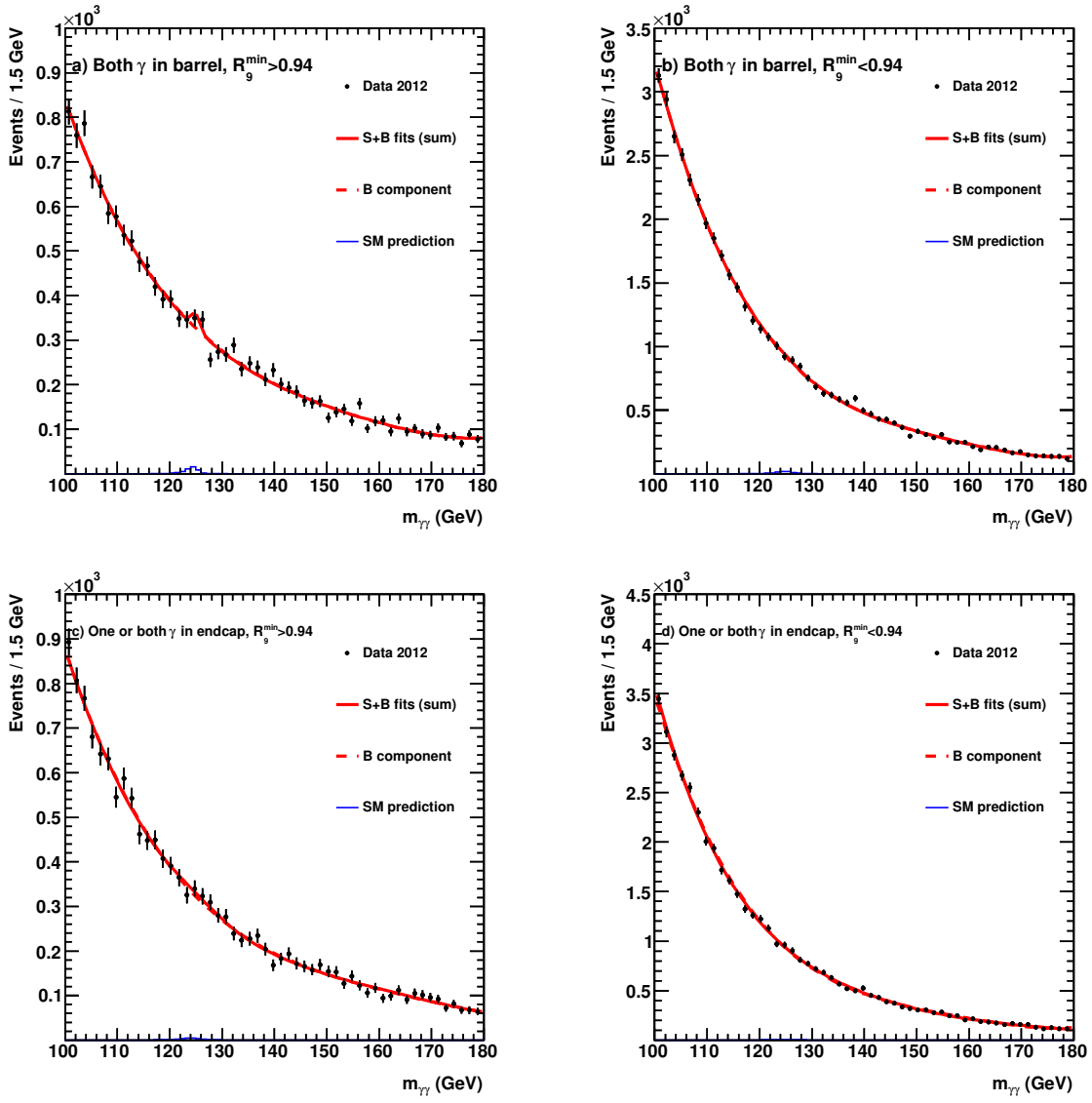


Figure 7: Mass distribution in individual classes for 2012 data with 5^{th} order polynomial plus Gaussian fit, with fixed a $\mu = 125GeV$ of the Gaussian.

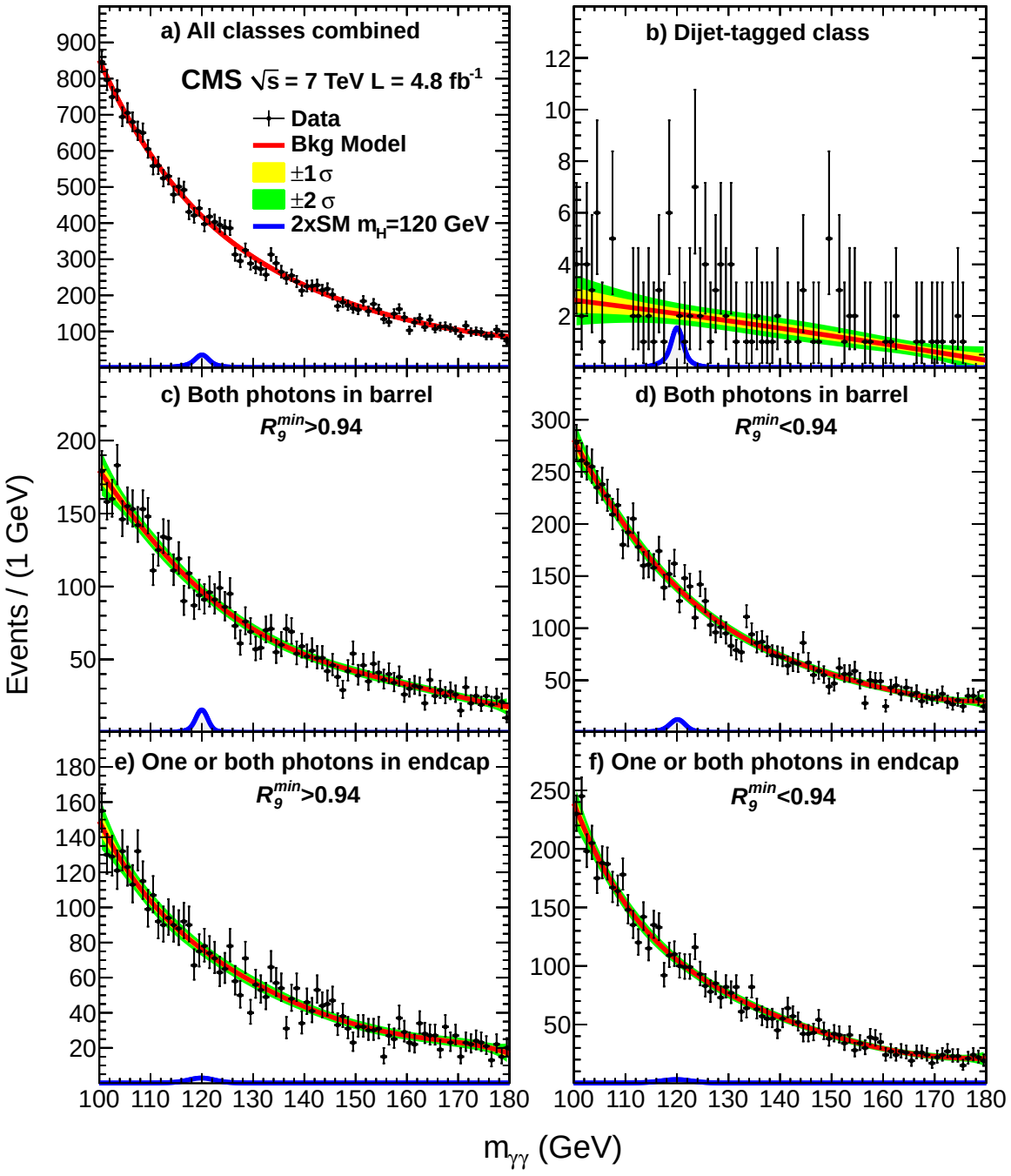


Figure 8: Paper mass distribution in individual classes for 2011 data taken from [4]

10.2 Trigger list

Table 1: List of triggers for 2012 data

Trigger name
HLT_Photon26_R9Id85_OR_CaloId10_Iso50_Photon18_R9Id85_OR_CaloId10_Iso50_Mass60
HLT_Photon26_R9Id85_OR_CaloId10_Iso50_Photon18_R9Id85_OR_CaloId10_Iso50_Mass70
HLT_Photon26_CaloId10_Iso50_Photon18_CaloId10_Iso50_Mass60
HLT_Photon26_CaloId10_Iso50_Photon18_R9Id85_Mass60
HLT_Photon26_R9Id85_Photon18_CaloId10_Iso50_Mass60
HLT_Photon26_R9Id85_Photon18_R9Id85_Mass60
HLT_Photon26_Photon18
HLT_Photon26_R9Id85_OR_CaloId10_Iso50_Photon18
HLT_Photon36_CaloId10_Iso50_Photon22_CaloId10_Iso50
HLT_Photon36_CaloId10_Iso50_Photon22_R9Id85
HLT_Photon36_R9Id85_OR_CaloId10_Iso50_Photon22_R9Id85_OR_CaloId10_Iso50
HLT_Photon36_R9Id85_Photon22_CaloId10_Iso50
HLT_Photon36_R9Id85_Photon22_R9Id85
HLT_Photon36_Photon22
HLT_Photon36_R9Id85_OR_CaloId10_Iso50_Photon22

10.3 Cut list

Table 2: List of cuts applied to 2011 data

Variable	Class 1	Class 2	Class 3	Class 4
rel. isolation sum	3.8	2.2	1.77	1.29
rel. track isolation	3.5	2.2	2.3	1.45
$\sigma^2_{i\eta i\eta}$	0.0106	0.0097	0.028	0.027
H/E	0.082	0.062	0.065	0.048
R_9	0.94	0.36	0.94	0.32

Table 3: Preselection of 2012 data on calorimeter and PF isolation

Variable	$R_9 < 0.9$	$R_9 \geq 0.9$
Corrected HCal Iso	4	50
Corrected Track Iso	4	50
Charged PFIso	4	4

Table 4: Cuts applied to 2012 data

Variable	Class 1	Class 2	Class 3	Class 4
PFIso chosen vertex	6	4.7	5.6	3.6
PFIso worst vertex	10	6.5	5.6	4.4
Charged PFIso	3.8	2.5	3.1	2.2
$\sigma^2_{i\eta i\eta}$	0.0108	0.0102	0.028	0.028
H/E	0.124	0.094	0.142	0.063
R_9	0.94	0.298	0.94	0.24

10.4 Technical description

This is the technical description to the analysis code. The description is also provided along with the code at GitHub[6] To run this analysis the usage of CernVM[7] is recommended, as this

[6]https://github.com/cms-opendata-analyses/2011-photon-2012-doublephoton-higgs-hgaga
[7]http://opendata.cern.ch/docs/cms-virtual-machine-2011

provides an environment ready for CMS analyses. If you have other resources to run CMS analyses, you can use these as well.
The next two sections explain how you can setup all needed tools and run the basic analysis. In the following sections we introduce you to the structure of the source code and how to modify specific parts of the analysis. We assume that you run the code on the CernVM.

10.4.1 Environment setup

First we need to setup a working area, where the CMS environment is setup and the code will be copied to.

```
mkdir WorkingArea
cd WorkingArea
cmsrel CMSSW_5_3_32
cd ./CMSSW_5_3_32/src
cmsenv
```

All tools used in the analysis are ready to use. We can clone the analysis source code from the GitHub repository and compile it.

```
git clone git://github.com/cms-opendata-analyses/2011-photon-2012-doublephoton-higgs-hgaga.git
scram b
```

As a last step of the setup, databases for accessing the datasets (AOD files) from CERN website are linked.

```
cd 2011-photon-2012-doublephoton-higgs-hgaga/Analyzer
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA FT_53_LV5_AN1
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/START53_LV6A1 START53_LV6A1
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT53_V21A_AN6_FULL FT53_V21A_AN6
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT53_V21A_AN6_FULL.db FT53_V21A_AN6_FULL.db
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT53_V21A_AN6_FULL FT53_V21A_AN6_FULL
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/START53_V27.db START53_V27.db
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/START53_V27 START53_V27
```

Now we are ready to start a run of the analysis.

10.4.2 Running the analysis

The analysis is split into two parts: **Analyzer** and **PostAnalyzer**.
First we need to convert the raw datasets (AOD files) from the CERN server to ntuples, which are stored locally. We provide a simple shell script (`2011-photon-2012-doublephoton-higgs-hgaga/Analyzer/run.sh`) to do so. Please take a look inside the script before you run it to get a general understanding of how it calls the analyzer. Before we run the analysis we compile the Analyzer again. Note that the shell script can be called with four different arguments to process different datasets or MonteCarlo simulations. Beware that if you run the analysis on the CernVM this process can take weeks to months. If you have a computer cluster available, which can handle CMS environment as setup above, you need to edit the shell script to submit jobs to your cluster. During the analyzer run some soft cuts are applied on the raw datasets to return events, which are interesting for further analysis.

```
cd 2011-photon-2012-doublephoton-higgs-hgaga/Analyzer
scram b
./run.sh 1
./run.sh 2
./run.sh 3
./run.sh 4
```

The analyzer creates `ROOT-ntuples` which we need to move to the PostAnalyzer directory for further analysis.

```
cd 2011-photon-2012-doublephoton-higgs-hgaga/
mv Analyzer/ntuples-data PostAnalyzer
mv Analyzer/ntuples-mc PostAnalyzer
```

To apply the cuts on the `ntuples`, plot corresponding mass distributions and a simplified significance test we provide three C++ scripts to run. The execution of these scripts should not take longer than two minutes (even on CernVM).

```
cd 2011-photon-2012-doublephoton-higgs-hgaga/PostAnalyzer
./compile.sh
./hggMakeHist
./hggMakePlots
./pvalPlot
```

This creates plots in the directory `2011-photon-2012-doublephoton-higgs-hgaga/PostAnalyzer/plots`, which can be compared to the results of the plots provided in the published results [4] and [8].
If you are interested in improving the analysis or use this as a template for other analysis purposes you can read through the following sections which give a deeper insight of the source code.

10.4.3 Structure

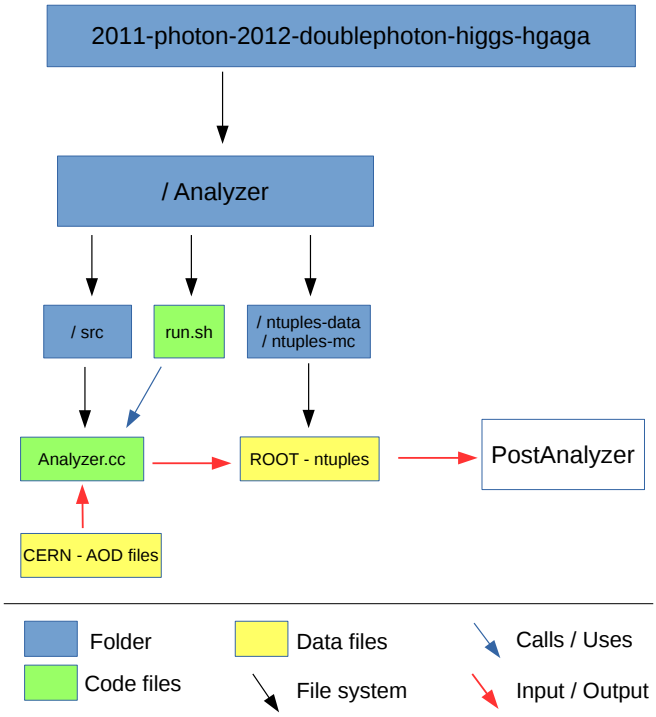


Figure 9: Data structure of `Analyzer`

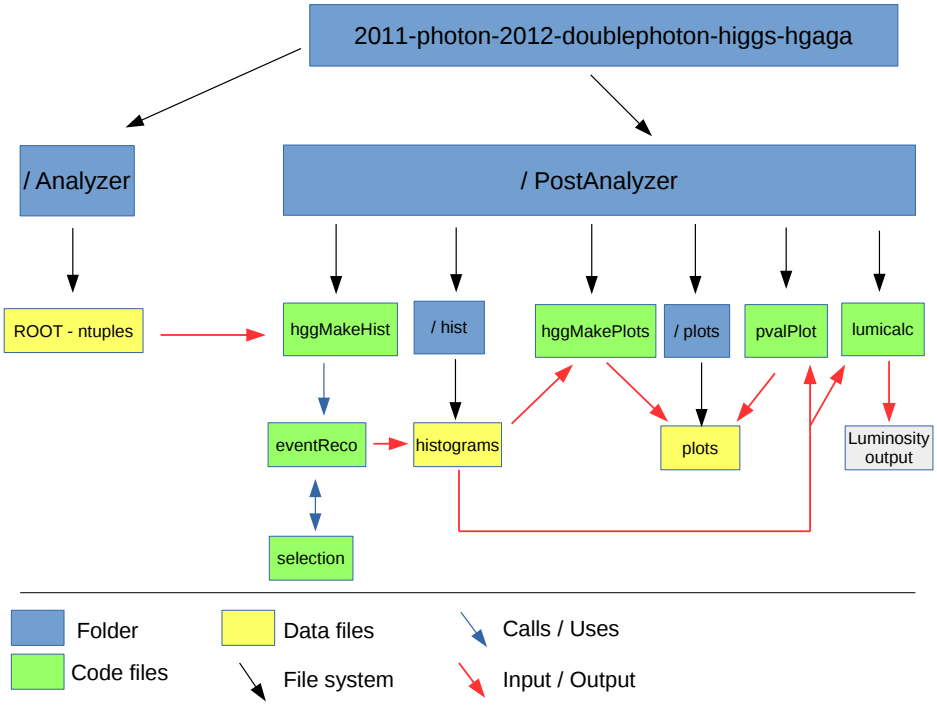


Figure 10: Data structure of `PostAnalyzer`.

The most important parts of the used data structure are summarized in figures 9 and 10. As stated before in the `Analyzer.cc` only soft precuts are applied to reduce the number of events in the `ntuples`. The cuts (in final version) are applied in the `selection.h`. `eventReco.h` delivers a framework for easier access and data storage of the histograms. Also it helps with analyzing data and MC - simulations. `hggMakePlots.cxx` uses the histograms and creates plots for the $m_{\gamma\gamma}$ mass distribution. These plots serve as an example and can be changed for several use cases. In the current state, these can be directly compared to plots from the papers.

10.4.4 Modifying files

If you want to edit the cuts, which we applied, the two interesting files are `Analyzer.cc` and `selection.h`.
In `Analyzer.cc`:
The file contains several functions, which are called at specific events. The most important function is `analyze()` which is called at each event. This function uses `SelectPhotons()`, which applies the soft-cuts to variables. Thus this is the main point to edit the physics part of this code. We use flags to refer to different dataset/mc-signals. These flags are tested in `if`-condition and then the corresponding soft cuts are applied.
In `selection.h`:
We have several functions in this file, which are called for the specific datasets. These functions are called within `SelectHgg()`. Cuts are applied in the functions and `histograms` are created from the `ntuples`.
The `histograms` are then used to create the result plots.
Important: After you edited the files you need to compile them again. For the `Analyzer.cc` that is done by the command `scram b` which can be called from anywhere inside `2011-photon-2012-doublephoton-higgs-hgaga/`. For the three scripts in `2011-photon-2012-doublephoton-PostAnalyzer/` this can be done by invoking `./compile.sh`. This script will also create all needed folders.

10.4.5 Adding more variables to the `ntuples`

To expand the analysis and use more variables these need to be added to the `ntuples`. As this requires a rerun of the `Analyzer`, variables should be chosen wisely beforehand.
In the analyzer a local variable/array is created for each variable in the `ntuples`. Afterwards this variable is added to a `ROOT - Tree`, which is then stored as the `ntuple`-file. For example below the implementation of the ratio of hadronic- to electromagnetic-energy ($\frac{H}{E}$) is shown in the `Analyzer.cc`.

```
// +++++ Analyzer.cc +++++
class Analyzer{
    //...
    private:
        //...
        //creating local array
        float _phHadronicOverEm[_maxNph];
        //...
}
Analyzer::Analyzer(...)
{
    //...
    //Add variable to tree branch
    _tree->Branch("phHadronicOverEm", _phHadronicOverEm, "phHadronicOverEm[Nph]/F");
    //...
}

int Analyzer::SelectPhotons(...)
{
    //...
    //Set the array values
    _phHadronicOverEm[_Nph] = it->hadronicOverEm();
    //...
}
```

To use this variables then in the `PostAnalyzer` we need to add them to the `Tree` which we use there. This is done by editing the file `2011-photon-2012-doublephoton-higgs-hgaga/PostAnalyzer/tree.h`. Here again an example for the $\frac{H}{E}$ variable is shown.

```
// +++++ tree.h +++++
class ZTree {
public:
    //...
    //create local variable
    Float_t          phHadronicOverEm[maxNph];
    //...
    //Add tree branch
    TBranch        *b_phHadronicOverEm;
    //...
}

void ZTree::Init(...)
{
    //...
    // Add local variable to tree
    fChain->SetBranchAddress("phHadronicOverEm", phHadronicOverEm, &b_phHadronicOverEm);
    //...
}
```

The variable is now available at the tree and can be used for example in the `selection.h` file to provide cuts.

```
//+++++ selection.h +++++
//...
double SelectPh11(...)
{
    //...
    // Example for accessing a tree variable
    // preselTree is a ZTree-pointer here
    if(preselTree->phHadronicOverEm[ph] > 0.082 && phClass == 3)
    //...
}
```

10.4.6 Luminosity calculation

To calculate the luminosity we provide an additional python script `PostAnalyzer/lumicalc.py`. This file needs the luminosity files `2011lumi.txt` and `2012lumi.txt`, which can be downloaded

from the Open Data Portal[8][9] and are also provided in the GitHub repository.
It is important to note that you need to **change the trigger selection** here, when you changed it in the Analyzer or PostAnalyzer.

References

[1] CMS Collaboration. Photon primary dataset in AOD format from RunA of 2011 (/Photon/Run2011A-12Oct2013-v1/AOD). *CERN Open Data Portal*, 2016. doi: 10.7483/OPENDATA.CMS.K3YX.WNFA.

[2] CMS Collaboration. DoublePhoton primary dataset in AOD format from Run of 2012 (/DoublePhoton/Run2012B-22Jan2013-v1/AOD). *CERN Open Data Portal*, 2017. doi: 10.7483/OPENDATA.CMS.CEPG.EXLP.

[3] CMS Collaboration. DoublePhoton primary dataset in AOD format from Run of 2012 (/DoublePhoton/Run2012C-22Jan2013-v2/AOD). *CERN Open Data Portal*, 2017. doi: 10.7483/OPENDATA.CMS.KT69.ANB8.

[4] CMS Collaboration. Search for the standard model Higgs boson decaying into two photons in pp collisions at $\sqrt{s} = 7$ TeV. *Physics Letters B*, 710(3):403–425, Apr 2012. doi: 10.1016/j.physletb.2012.03.003.

[5] CMS Collaboration. Search for a Higgs boson decaying into two photons in pp collisions recorded by the CMS detector at the LHC. *CMS Internal*, Aug 2012.

[6] CMS Collaboration. Search for the Standard Model Higgs boson decaying into two photons. *CMS Internal*, Aug 2012.

[7] CMS Collaboration. Updated measurements of the new Higgs-like boson at 125 GeV in the two photon decay channel. *CMS Internal*, Jan 2013.

[8] CMS Collaboration. Observation of the diphoton decay of the Higgs boson and measurement of its properties. *The European Physical Journal C*, 74:3076, Oct 2014. doi: 10.1140/epjc/s10052-014-3076-z.

[9] CMS Collaboration. Simulated dataset GluGluToHToGG_M-125_8TeV-powheg-pythia6 in AODSIM format for 2012 collision data. *CERN Open Data Portal*, 2017. doi: 10.7483/OPENDATA.CMS.BSFG.VPS2.

[10] CMS Collaboration et al. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004–S08004, Aug 2008. doi: 10.1088/1748-0221/3/08/s08004.

[8] http://opendata.cern.ch/record/1051
[9] http://opendata.cern.ch/record/1052