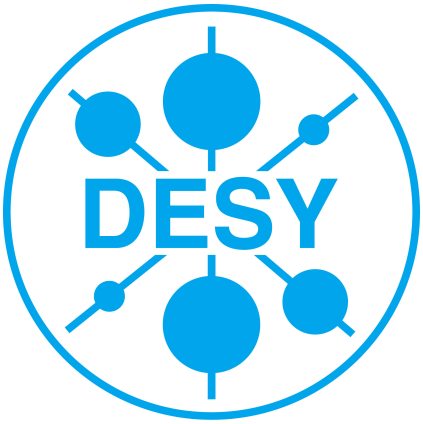
CMS Open Data Project

Rediscovery of long-range azimuthal angular correlations for CMS Open data p-p collisions at $\sqrt{S_{NN}} = 7$ TeV

Adrian Nassirpour
Supervised by: Achim Geiser
Lund University, Sweden
Deutsches Elektronen-Synchrotron (DESY), Hamburg

September 7, 2016



Abstract

Results on long-range two-particle azimuthal correlations for charged particles in CMS Open Data proton-proton (p-p) collisions at $\sqrt{S_{NN}} = 7$ TeV are presented, using Open Data from the 2nd half of 2010, over a full azimuthal coverage & a pseudorapidity coverage of $0.1< \eta <2.4$. The two-particle correlation function is derived from a similar, official CMS study written in late September, 2010, using the 1st half of the 2010 p-p, $\sqrt{S_{NN}} = 7$ TeV, data [1]. The results from the CMS Open Data study are then directly compared to the results from the original 2010 study. The results between the two studies are found to be qualitatively similar; just as in the original study, a ridge-like structure emerges across $2< \Delta\eta <4$ & $\Delta\phi \approx 0$ when observing the 2-particle correlation function for high-multiplicity data. Therefore this project further helps to validate the CMS 2010 Open Data set.

Acknowledgments

I wish to thank the DESY Summer Student Program for the wonderful opportunity to perform this study. A special thanks to my supervisor Achim Geiser for not only giving me an exciting project, but also managing to make the whole experience fun and encouraging me to pull through. Additional thanks go to the patient people that I had to share offices with, Oleksandr Kot & Nazar Stefaniuk, for help during dire times, and for putting up with my constant banter with Achim. Lastly, a thanks to Olaf Behnke for organizing various trips & activities throughout the student program. I had a great overall experience at DESY, and I would encourage other students to apply for the program in the future!

1 Introduction

The main goal of this project is to try and recreate the Compact Muon Solenoid (CMS) study in long-range azimuthal correlations [1], in order to further validate the CMS 2010 p-p, $\sqrt{S_{NN}} = 7$ TeV, Open Data set. By doing so, this analysis will also showcase the benefits of the open data project; that physicists outside of the CMS collaboration, or even non-physicists, can run their own analyses or studies on data gathered from CMS. This can in turn open new possibilities for research by cross-referencing studies & analyses between different experiments.

Throughout this paper, the specifics of the two-charged particle correlation function, used in both the original 2010 study and this study on CMS 2010 Open Data, will be discussed, as well as any discrepancies between the two studies. This section, more specifically, will start discussing some basic High-Energy Physics (HEP) phenomenology, as well as a brief overview of the CMS detector.

1.1 CMS Open Data Project

The CMS collaboration is the first among the 4 main experiments at the Large Hadron Collider (LHC) to release a part of its data to the open public [2]. The main goal of the CMS Open Data project is to provide the data, alongside similar analytic tools used by the CMS, to the open public. By doing so, one should be able to analyze data from CMS in a "home-like" environment. The data that is open to the public is unselected, but calibrated, research-level data. Data where no post-detector cuts have been applied, and where detector triggers can be manipulated freely. While there are only a quite limited amount of documentation open to the public at the time of writing this report, the CMS Open Data project has come far since its inception.

1.2 Brief Overview of Basic HEP Phenomenology

1.2.1 Transverse Momentum

One of the key variables to observe during a high-energy physics experiment is the momentum, and how it is conserved, of tracks and other objects throughout the system. The CMS detector is, as well as most particle detectors are, built concentrically around the particle beam-axis. Thus, the momenta of the particles that scatter perpendicular to the beam-axis are easier to measure, and defined as the transverse momentum, p_{T} ,of the corresponding particle. The transverse momentum is referred to as:

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} \tag{1}$$

Where p_x and p_y are the Cartesian values of the momentum perpendicular to the beam-axis. Furthermore, the p_{T} for the incoming protons is zero before the collision (the total momentum is focused along the beam axis), and non-zero for the individual outgoing particles. Observing the p_{T} for each particle is therefore giving a picture of what occurs during the collision.

1.2.2 Azimuthal and Polar angles

For cylindrical coordinates, two angles are measured for each particle track. The polar angle, also known as the scattering angle, θ, is the longitudinal angle between the beam-line and the particle track. In Cartesian coordinates, assuming the z-direction is along the beam-line, it represents the angle from the z-axis to the particle track.

In contrast, the azimuthal angle ϕ is defined as the horizontal angle between a track and an arbitrary axis, around the beam-line. In Cartesian coordinates, assuming that the z-direction is along the beam-line, it represents the angle from the x-axis to the particle track.

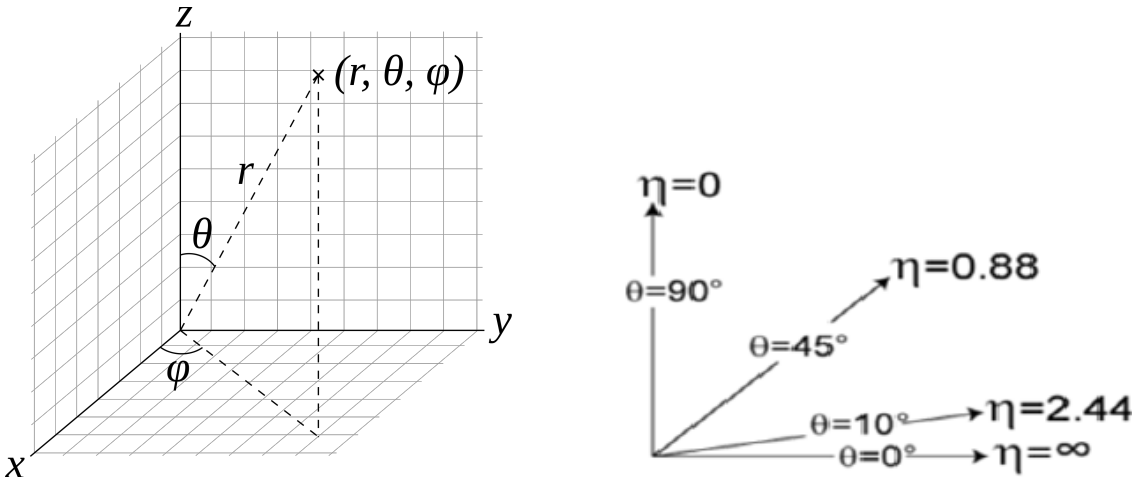


Figure 1: Figures illustrating common HEP variables. The left-hand figure illustrates the azimuthal and polar angles for spherical coordinates [3] . The right-hand figure showcases the relation between the scattering angle and pseudo-rapidity [4].

1.2.3 Rapidity and Pseudo-rapidity

In order to measure the linear components of the momentum, two new factors are introduced. Namely the rapidity and the pseudo-rapidity. The rapidity of the detector system is defined as[X]:

$$y = \ln \frac{E + p_z}{E - p_z} = \tanh^{-1} v_z \tag{2}$$

Where E is the energy of the particle ($E = \sqrt{m^2 + \boldsymbol{p}^2}$, where $\boldsymbol{p}$ is the total momentum), p the different momentum components and v_z is the parallel velocity. As the particles in the accelerator move at relativistic speeds, it is thus more useful to measure the rapidity instead of the standard velocity. Furthermore, rapidity is Lorentz additive for boosts along the beam-axis. However, rapidity can sometimes be hard to measure, therefore an approximation of the rapidity, namely pseudo-rapidity, is used instead. The pseudorapidity is defined as:

$$\eta = \frac{1}{2} \ln \frac{|\boldsymbol{p}| + p_z}{|\boldsymbol{p}| - p_z} = -\ln \tan(\frac{\theta}{2}) \tag{3}$$

Where $\boldsymbol{p}$ and p_z are defined like the rapidity variables, and θ is the scattering angle. Since the rapidity is Lorentz additive for boosts along the beam axis, assuming that pseudorapidity is a sufficient approximation of rapidity, then pseudorapidity will also be Lorentz additive. As is seen in equation 3, in the limit of mass-less particles (such as photons), pseudo-rapidity becomes identical to rapidity.

1.3 Brief Overview of the CMS Detector

The CMS detector is one of the 4 large detectors that operate at the LHC. The detector is set up concentrically around the beam-line, with several onion-shaped layers of different detectors. As the name suggest, the detector contains a large solenoid magnet, with a diameter of 7m, boasting a magnetic field up to 3.8T[5]. The detector has full azimuthal coverage, as well as a pseudorapidity coverage of $0.1 < \eta < 2.5$ for the main tracking chamber [6].

The detector layout can be seen in fig.2, and illustrates the different detector layers and their functions. The components that are important for this study are enclosed within the solenoid magnet; mainly the silicon trackers. The silicon trackers do not only detect the spatial coordinates of each track, but can also be used in order to calculate their p_T. This is done by measuring the bending radius of each track due to the applied magnetic field[7].

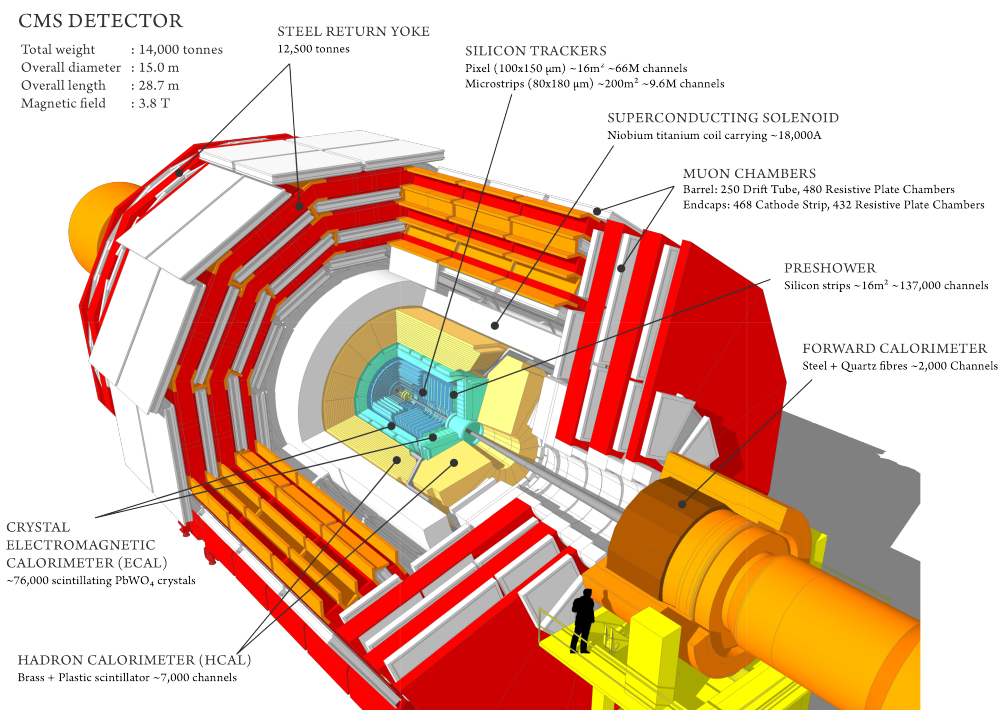


Figure 2: An illustration of the CMS detector. and its various components [5]

2 Discovery of Ridge from 2010 study

This report will mainly focus on validating and recreating the discovery of long-range angular correlations for p-p data. The first occurrence of this phenomenon was published in 2010[1], using the first half of the 2010 data. Henceforth, any mentions in this report of "the 2010 study/analysis" or "the original study" will be in reference to this paper. The 2010 study is divided into two analyses; one focused on short-range angular correlations ($|\Delta\eta| < 2$), and one focused on long-range angular correlations ($2.0 < |\Delta\eta| < 4.8$).

The data used in the 2010 analysis is what is described as minimum bias data; data that is taken with as little detector bias as possible. This is done by obtaining the events that do not strike a main trigger, but instead strike a auxiliary trigger. These auxiliary triggers contain the minimal definition of whether a track exists or not, thus the data is called minimum bias. This is not the be confused with zero bias data, which is defined by sampling all the data that does not strike the main triggers. In contrast to minimum bias, zero bias data leaves room for recording "empty" events, whilst minimum bias still retains a certain threshold.

In addition to the minimum bias data, two other high multiplicity pixel triggers (HLT) were used in order to gather high multiplicity data. These triggers play an important role in gathering extra high multiplicity data for the investigation of the long-range correlations, as the "ridge phenomena" is only apparent at multiplicities over 110.

The main result, where the long-range angular correlations can be seen, is found below in fig.3. The analytically functions displayed in the figure will be discussed more thoroughly in Section 3. Comparing the low-multiplicity case against the high-multiplicity plot, one can note that they both have ridge-like structures around $|\Delta\eta| \approx \Delta\phi \approx 0$, as well as one large ridge at $\Delta\eta \approx \pi$ across the whole $\Delta\eta$ range. However, both of these ridges were expected at the time. The novel, unexpected, feature here is the small ridge around $\Delta\phi \approx 0$, $2 < |\Delta\eta| < 4$, which can only be seen for the high-multiplicity case. Until this study, this phenomena was only expected to occur in heavy-ion collisions [8]. Recreating this small ridge will be the main goal throughout this CMS Open Data analysis.

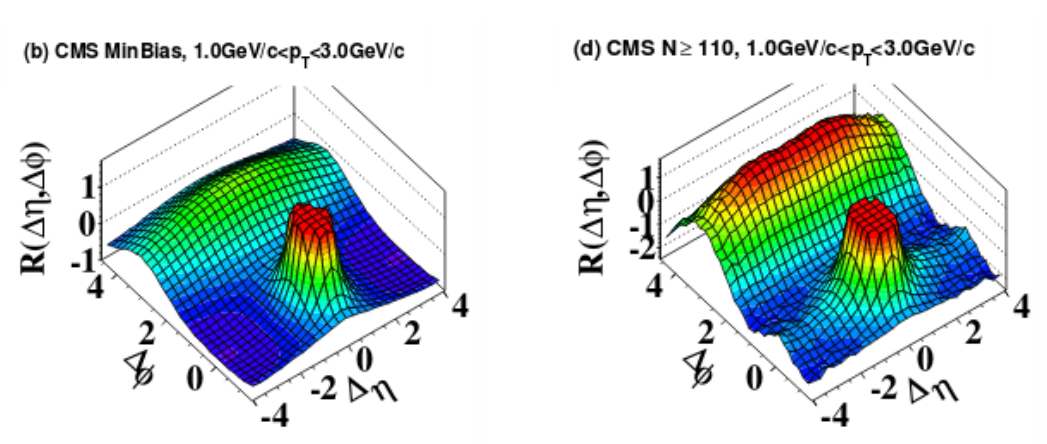


Figure 3: Results showcasing the 2-particle correlation function from the 2010 study[1]. The left-hand figure displays the 2-particle correlation function for inclusive MinBias data, with a kinematic 1.0GeV/c< p_T <3.0GeV/c cut. The right-hand figure displays the same correlation function and kinematic cut, but for events with a multiplicity over 110, using all triggers.

3 Method of Analysis

3.1 Short-Range & Long-Range Correlation Analysis

In order to see whether this analysis was viable or not, the data distribution of the Open Data set had to be compared with the data distribution from the 2010 study. This was done by selecting the appropriate minimum bias triggers, along with the HLT70, HLT85 and HLT100 triggers used in the 2010 experiment. After applying the appropriate triggers, a kinematic cut of $|\eta| < 2.4$, $p_T > 0.4$GeV/c was applied to the tracks associated to the "best" vertex (the vertex containing the largest amount of connected tracks), to ensure a low fake-rate. The remaining multiplicity for each such vertex is in this case defined as $N_{trk}^{offline}$. The association between tracks and verticies was simply done by backwards-extrapolating the tracks back to a common starting point (with appropriate uncertainties; see track collection used [9]) where they all intersect. The vertex definitions differs somewhat between the CMS Open Data study and the 2010 study; these differences will be covered in section 4.2.

Once the total data distribution was reconstructed, the events were divided into 10 + 4 bins in track multiplicity, which in this case equals to $N_{trk}^{offline}$. The 10 bins were used in order to track $N_{trk}^{offline}$ from 3 to ∞, in intervals that can be seen below. This binning was mainly used for the short-range correlations. Each bin roughly contained the same number of average entries:

- 10-bin 7 TeV: [3,7],[8,9],[10,11],[12,13],[14,17],[18,23],[24,29],[30,37],[38,49],[50,∞]

In contrast, the 4 other bins were used for the high-multiplicity comparisons, and are used in order to study long-range correlations. The binning is defined as follows:

- 4-bin 7 TeV: [3,34],[35,89],[90,109],[110,∞]

The reason for having 10+4 bins is simply to reconstruct the 2010 paper as accurate as possible. As is hinted in the paper, but more so confirmed in its corresponding analysis note [10], is that the short-range correlation 10-bin system was the original plan for the project. However, as the ridge effect was only seen in high-multiplicity data, they quickly implemented the 4-bin system, but chose to still present some results using the 10-bin system. The 10 bins in this report are the same as the ones in the analysis note.

The charged two-particle correlation analysis was first individually done within each bin, and then averaged together in order to arrive at the raw two-particle correlation function. The correlation function itself is dependent on two analytic functions; one for the signal distribution S_{NN},

and one for the background distribution B_{NN}. S_{NN}, which is merely the charged two-particle pair density function, is defined as follows:

$$S_{NN}(\Delta\eta, \Delta\phi) = \frac{1}{N_{event}^{signal}(N_{event}^{signal} - 1)} \frac{d^2N^{signal}}{d\Delta\eta d\Delta\phi} \tag{4}$$

where N_{event}^{signal} corresponds to the number of tracks processed in the signal distribution for each event. This number may differ from $N_{trk}^{offline}$, as additional kinematic cuts can be applied after the event has been slotted into its appropriate $N_{trk}^{offline}$ bin. These cuts will be displayed on the relevant figures in Section 4.1. If nothing else is mentioned for a specific plot, the same kinematic $|\eta| < 2.4$, $p_T > 0.4$GeV/c cut from $N_{trk}^{offline}$ was also applied to N_{event}^{signal}.

$\frac{d^2N^{signal}}{d\Delta\eta d\Delta\phi}$ is defined as a tracks $\Delta\eta$ ($\Delta\eta = \eta_1 - \eta_2$) and $\Delta\phi$ ($\Delta\phi = \phi_1 - \phi_2$) as compared to any other given track in the event. This quantity is recorded for every given pair in the event, and is then weighted by the total number of pairs ($N_{event}^{signal}(N_{event}^{signal} - 1)$) in said event, event-by-event. The distribution averaged between all the different $N_{trk}^{offline}$ bins, and finally normalized to the unit integral.

In contrast, the background distribution B_{NN}, which is represents a collection of completely uncorrelated particle pairs, is defined as follows:

$$B_{NN}(\Delta\eta, \Delta\phi) = \frac{1}{N_{event}^{signal} N_{prv.event}^{bkgrd}} \frac{d^2N^{mixed}}{d\Delta\eta d\Delta\phi} \tag{5}$$

here, $N_{prv.event}^{bkgrd}$ is simply defined by the N_{event}^{signal} from the previous event, in the same multiplicity bin. $\frac{d^2N^{mixed}}{d\Delta\eta d\Delta\phi}$ is defined as a tracks $\Delta\eta$ and $\Delta\phi$, in the current event (where the signal distribution is collected), compared to any other given track across another event. This quantity is recorded for every given pair across the two events, and then weighted by the total number of pairs across the two events $N_{event}^{signal} N_{prv.event}^{bkgrd}$, event-by-event. This will ensure a mixed distribution that is completely uncorrelated, as $\Delta\eta$ and $\Delta\phi$ are calculated between two completely different events. Lastly, just as with the signal distribution, the background distribution is averaged between all the different $N_{trk}^{offline}$, and finally normalized to the unit integral. The background distribution is calculated with mixed events within each $N_{trk}^{offline}$ multiplicity bin, in order to ensure that there is sufficient background statistics for high multiplicity events. This is due to the fact that low-multiplicity events have a much higher frequency than high-multiplicity events. Thus, if one would always calculate the background by mixing the current even with the "previous" event, there is a great possibility that the $N_{trk}^{offline}$ >110 event is mixed with a 3< $N_{trk}^{offline}$ <7 event.

For both the background and signal, the full $\Delta\phi$-range was stored, but only the absolute $\Delta\eta$ range was stored. This is due to the fact that the tracks order in each event should not make any physical difference. For aesthetic purposes, the $\Delta\eta$ range is then mirrored across the negative values. Furthermore, an additional kinematic cut, $|\Delta\eta| > 0.06, |\Delta\phi| > 0.06$, is applied in order remove track pairs from photon conversions, and other residual effects.

Once the signal and background distributions are reconstructed, the charged two-particle correlation function $R(\Delta\eta, \Delta\phi)$ can be calculated:

$$R(\Delta\eta, \Delta\phi) = \left\langle N_{avrg} \left(\frac{S_{NN}(\Delta\eta, \Delta\phi)}{B_{NN}(\Delta\eta, \Delta\phi)} - 1 \right) \right\rangle_N \tag{6}$$

The track multiplicity factor N_{avrg} (also referred to as N-1 in the 2010 study) represents the average multiplicity in each multiplicity bin, whereas $\frac{S_{NN}}{B_{NN}} - 1$ simply translates to $\frac{S_{NN} - B_{NN}}{B_{NN}}$, which represents the ratio between the signal and background distribution. As B_{NN} is completely uncorrelated, the ratio between S_{NN} and B_{NN} will be able to display if there are any correlations within S_{NN}. The projections for both $\Delta\eta$ and $\Delta\phi$ were recorded for the correlation functions ($R(\Delta\eta), R(\Delta\phi)$), as well as for the background and signal distributions respectively. They are each respectively defined as:

$$R(\Delta\eta) = \left\langle N_{avrg} \left(\frac{S_{NN}(\Delta\eta)}{B_{NN}(\Delta\eta)} - 1 \right) \right\rangle_N \tag{7}$$

$$S_{NN}(\Delta\eta) = \int S_{NN}(\Delta\eta, \Delta\phi) d\Delta\eta \tag{8}$$

$$B_{NN}(\Delta\eta) = \int B_{NN}(\Delta\eta, \Delta\phi) d\Delta\eta \tag{9}$$

$$R(\Delta\phi) = \left\langle N_{avrg} \left(\frac{S_{NN}(\Delta\phi)}{B_{NN}(\Delta\phi)} - 1 \right) \right\rangle_N \tag{10}$$

$$S_{NN}(\Delta\phi) = \int S_{NN}(\Delta\eta, \Delta\phi) d\Delta\phi \tag{11}$$

$$B_{NN}(\Delta\phi) = \int B_{NN}(\Delta\eta, \Delta\phi) d\Delta\phi \tag{12}$$

3.2 Recap of Various "N-Labeled" Variables

There are several variables labeled as "N" throughout the study. Below is a table giving a brief recap of their definitions: Observing Table 1, it is important to note that the 4 bottom multiplicity

Table 1: A table containing the various "N-like" variables

Function	Variable	Definition
n/a	$N_{trk}^{offline}$	The number of tracks connected to the best vertex alongside a $\|\eta\| < 2.4$, $p_T > 0.4$GeV/c kinematic cut
$S_{NN}(\Delta\eta, \Delta\phi)$	$N_{event}^{signal}(N_{event}^{signal} - 1)$	The number of pairs processed for the signal distribution. May contain additional cuts compared to $N_{trk}^{offline}$
$B_{NN}(\Delta\eta, \Delta\phi)$	$N_{event}^{signal} N_{prv.event}^{bkgrd}$	The number of pairs processed for the background distribution. Simply mixes N_{event}^{signal} from two different events.
$R(\Delta\eta, \Delta\phi)$	N_{avrg}	The average multiplicity over all signal events
$R(\Delta\eta, \Delta\phi)$	N	The average multiplicity over all multiplicity bins

variables do not necessarily have to coincide with $N_{trk}^{offline}$. Once an event has been put into its appropriate $N_{trk}^{offline}$ bin, further kinematic p_T cuts may be applied. Thus the number of tracks that play a part in the correlation analysis may be different the number of $N_{trk}^{offline}$ tracks it contains. Furthermore, it may be wise to divide $N_{event}^{signal}(N_{event}^{signal} - 1)$ with a factor of 2 while doing the signal analysis, as it is unnecessary to fill the same pair correlation twice (track 1 & track 2 compared to track 2 & track 1 etc). This is not explicitly stated in the 2010 study, but was assumed to remove additional redundancy. However, the same does not hold true for the background distribution; the more data yielded for the background, the better. The additional factor of 2 (or, any other factor depending on how much background one wants) will still cancel out compared to the signal distribution, whilst normalizing to the unit integral for both distributions.

4 Results & Discussion

4.1 Preliminary Results

As explained in section 4.1, the first necessary cornerstone of this project was to ensure that the inclusive data distribution of the CMS 2010 Open p-p data was similar to the data used in the 2010 study. The comparison between the two data-sets can be seen below in fig.4

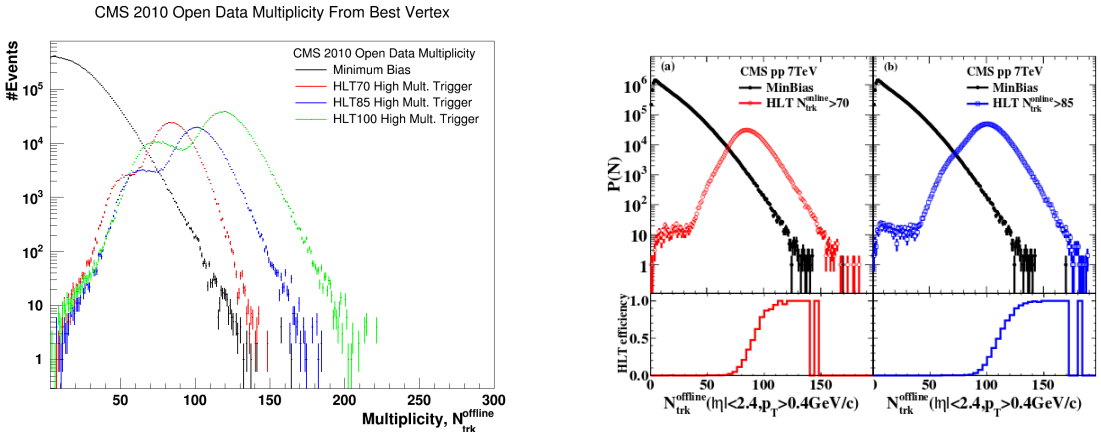


Figure 4: Inclusive data distributions of $N_{trk}^{offline}$ used in analysis, with their respective triggers. The left-hand figure illustrates the data distribution used in the CMS Open Data analysis. The right-hand figure illustrates the data distribution used in the original 2010 study[x].

As is seen in fig.4, the data distributions between the two studies are qualitatively similar. The exception of this is the "left-shoulder" seen for the HLT-triggers in the left-hand part of fig.4. This behavior is probably due to the different vertex definition of $N_{trk}^{offline}$ between the two studies, which will be discussed in more detail in section 4.2. All things considered, the bump of low-multiplicity events from the HLT-triggers do not play any role of importance to the final analysis. This is due to the fact that the final results mostly consists of comparing the Minimum Bias trigger with events that have an $N_{trk}^{offline} > 110$, thus circumventing this bump entirely.

Table 2: Number of $N_{trk}^{offline}$ events for different triggers and bins. The negative numbers in the third column represents a deficit, of the same bin, compared to the 2010 study.

$N_{trk}^{offline}$ & Trigger	Event Count	Diff. to 2010
Inclusive MinBias	9.568M	-11.862M
$N_{trk}^{offline}$ >110, MinBias+HLT70+HLT85	109.6K	-244.4K
$N_{trk}^{offline}$ >100, MinBias+HLT70+HLT85+HLT100	904.3K	550.3K

Table 2 displays a breakdown of the numbers for the entries in fig. 2. Comparing these numbers to the same distribution in the original paper, one can find that the total number of MinBias events in the Open Data study is less than one half of the original study. Using only the same HLT triggers as in the original study for the high-multiplicity events, the CMS Open Data study only yields a third of the original distribution. However, when applying the HLT100 trigger, the total number of high-multiplicity events exceed the original paper by almost a factor of 3.

Moving onwards, the data from fig.4 was put into their appropriate multiplicity bins as explained in section 4.1, to then have the signal & background distributions calculated accordingly to equations 4 & 5 respectively. The only $N_{trk}^{offline}$ bin presented in the results is the $N_{trk}^{offline}$ >110 bin. This is due to the fact that the long-range correlation ridge is most prevalent in this bin.

Since there is no Monte-Carlo simulation for the CMS 2010 Open Data to compare with, showing the individual bins made little sense. This of course will be made more prevalent if the same analysis is applied unto CMS 2011 Open Data.

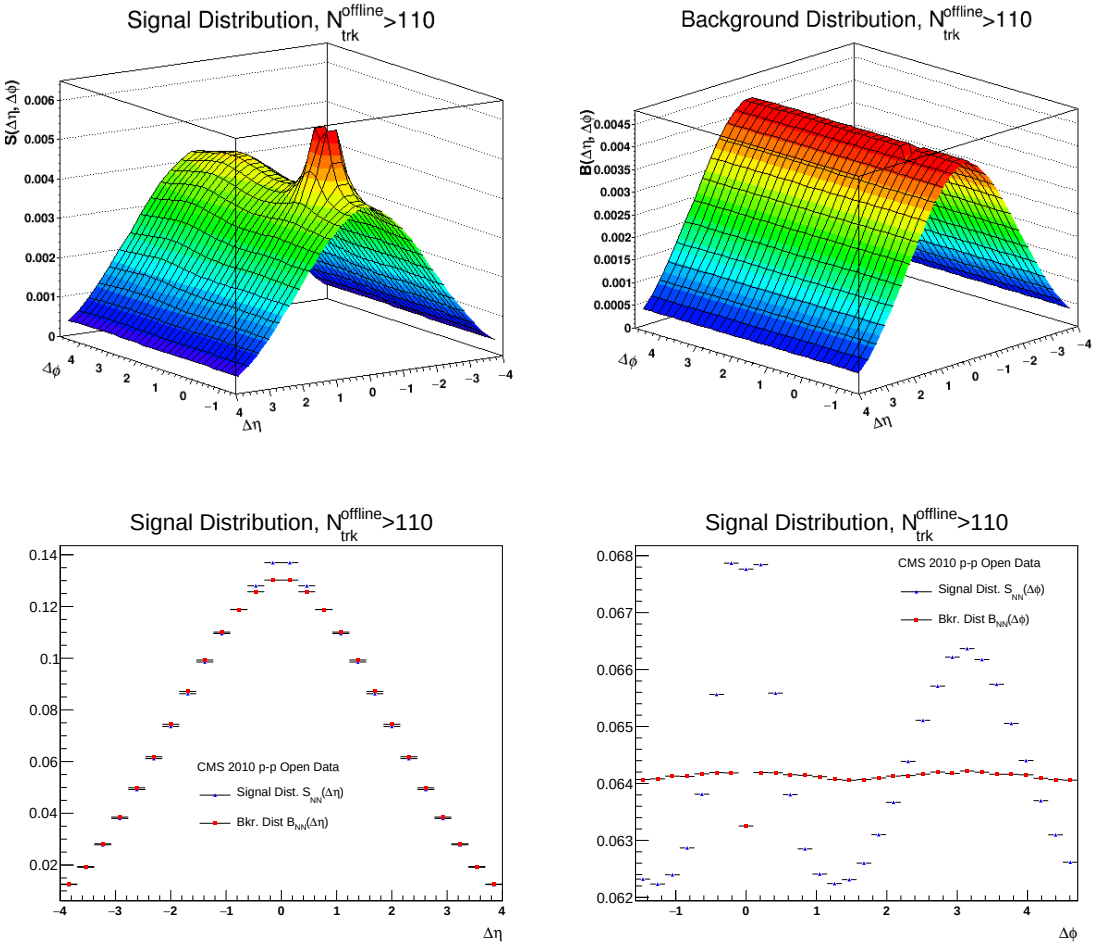


Figure 5: Several plots showcasing the signal and background distributions for the 2-particle correlation for the $N_{trk}^{offline}$ >110 bin . a) The total signal distribution $S_{NN}(\Delta\phi, \Delta\eta)$ for $N_{trk}^{offline}$. b) The total background distribution $B_{NN}(\Delta\phi, \Delta\eta)$ for $N_{trk}^{offline}$. c) The $\Delta\eta$ projection from $B_{NN}(\Delta\phi, \Delta\eta)$ & $S_{NN}(\Delta\phi, \Delta\eta)$ respectively. d)The $\Delta\phi$ projection from $B_{NN}(\Delta\phi, \Delta\eta)$ & $S_{NN}(\Delta\phi, \Delta\eta)$ respectively.

As can be seen through fig's 5a-d, there is a stark contrast between the signal and background distributions. Both distributions are normalized to the unit integral. Using the 2-particle correlation function as described in equation 6, as well as applying a 1.0GeV/c< p_T <3.0GeV/c kinematic cut on both particles in each correlation pair(comparatively to fig.x in section 3), the final 2-D two-particle correlation distribution for the $N_{trk}^{offline} > 110$ can be obtained, and is displayed in fig.6.

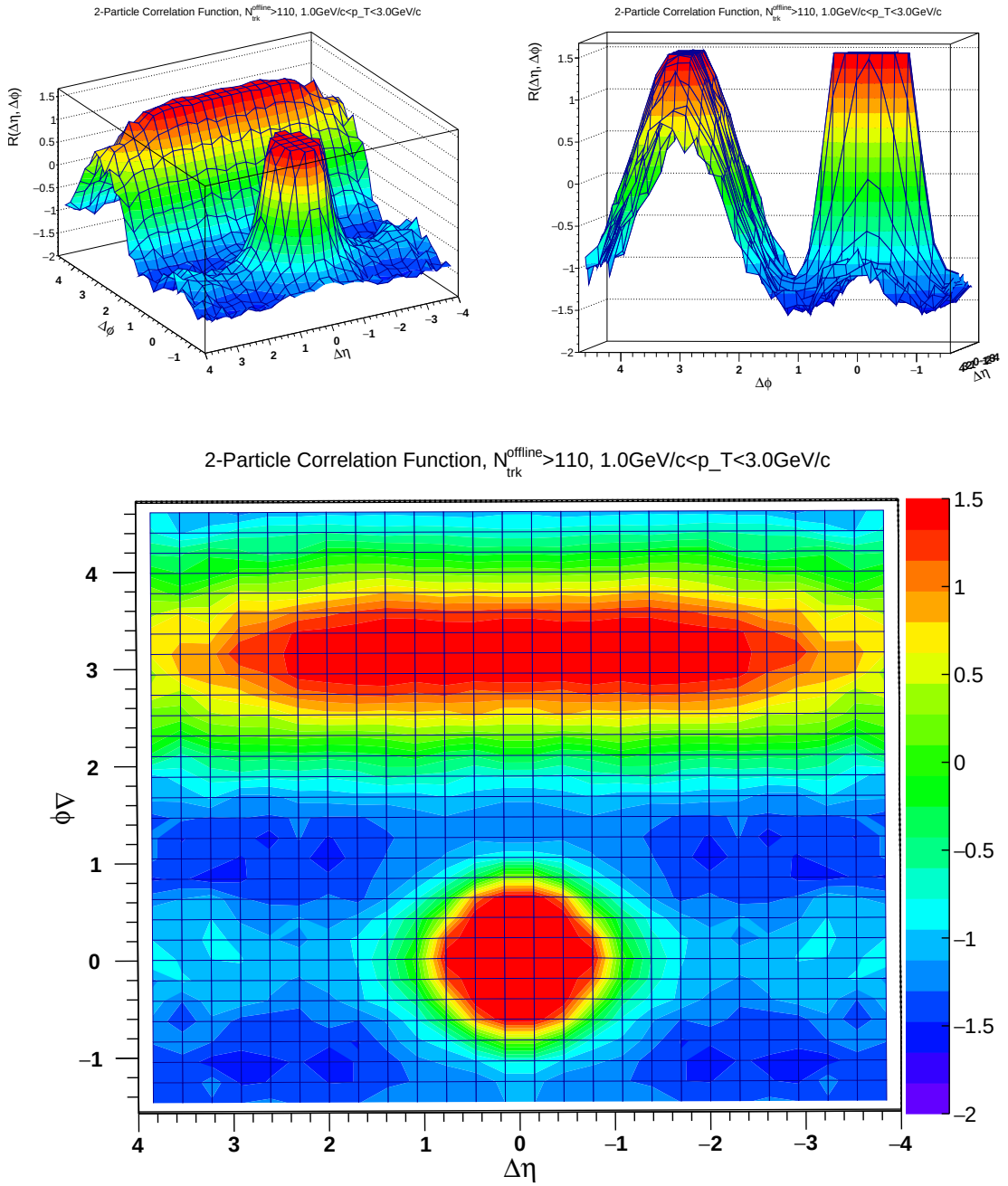


Figure 6: The 2-particle correlation function $R(\Delta\eta, \Delta\phi)$, for $N_{trk}^{offline}$ >110 with a kinematic cut of 1.0GeV/c< p_T <3.0GeV/c, as seen from three different angles. The main-leading jet peak around $\Delta\eta \approx \Delta\phi \approx 0$ is severely cut-off in order to exhibit the structure outside of that region.

4.2 Diversions From the 2010 Study

There were several diversions & differences done in this project as compared to the original 2010 discovery, thus a fully quantitative comparison is not possible. Some of the differences will be covered in this section.

Firstly, the data set from the CMS Open Data project differs from the ones used in the original 2010 project. The data set used in the 2010 project can be found here[10], and consists of minimum

bias data from the first half of the 2010 7 TeV run. In contrast, the data sets used in this analysis can be found here[9], and consists of minimum bias data from the second half of the 7 TeV run.

The original project contains a slightly different definition of what to include in $N_{trk}^{offline}$. Instead of simply including the tracks connected to the largest vertex for any given event, an auxiliary condition is added to include tracks that have a small separation between the beam axis,z, and the impact parameter,b, relative to the primary vertex. Furthermore, in addition to $N_{trk}^{offline}$ multiplicity bins, the original study contained vertex bins. Thus the correlation function $R(\Delta\eta, \Delta\phi)$ was only calculated for signal and background events that had their primary verticies in approximately the same spots inside the detector. Both of these definitions were omitted due to practical reasons, but would have been included if no ridge-like structure would have been seen without them. Fortunately, this was not the case.

Moreover, as opposed to the data used in the 2010 study, the CMS 2010 Open Data data set does not contain Monte-Carlo (MC) simulations. Whilst the CMS 2011 Open Data set does include MC simulations, the triggers from the 2010 Open Data set do not necessarily directly translate to the 2011 Open Data triggers. Thus further investigation into the triggers would be required before implementing the same study to the 2011 CMS Open Data sets.

5 Conclusion

The 2010 study in long-range angular correlations, using the charged 2-particle correlation function, for 7 TeV p-p collisions has been qualitatively reconstructed, thereby validating the CMS 2010 7 TeV p-p Open Data. Observing fig.x, the correlation function was reconstructed in such a way that the small ridge is apparent in the region of $\Delta\phi \approx 0$, $2 < \Delta\eta < 4$, suggesting long-range angular correlations. Comparing fig.x with the same plot from the original study, even though the two studies diverge in certain aspects, one can find that they retain the same qualitative aspects. Thus validating that the CMS 2010 Open Data is indeed at "research-level" caliber.

Looking forwards, what could be done is to apply the same study on CMS 2011 7 TeV p-p Open Data. As the 2011 sample contains MC simulations, a more quantitative approach to the original 2010 study can be performed.

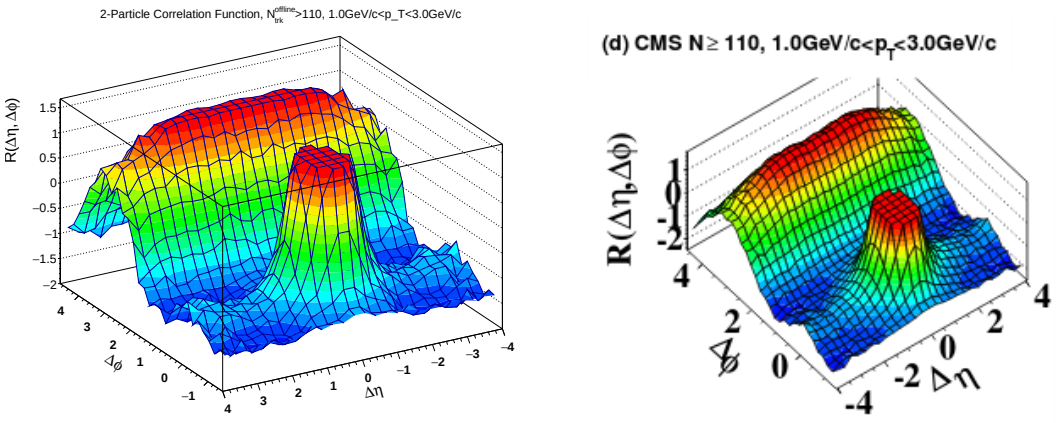


Figure 7: A final figure directly comparing the results from the CMS 2010 Open Data analysis (left), to the results from the original 2010 study (right)

6 References

[1] "Observation of Long-Range Near-Side Angular Correlations in Proton-Proton Collisions at the LHC", the CMS Collaboration, https://arxiv.org/abs/1009.4122. *Last visited:* 07/09-2016

[2] CMS Open Data Project http://opendata.cern.ch/research/CMS *Last visited:* 07/09-2016

[3] Figure showing azimuthal angle
https://commons.wikimedia.org/wiki/File:3D_Spherical.svg. *Last visited:* 07/09-2016

[4] Figure showcasing pseudorapidity, "Neural Networks Search for Charged Higgs Boson of Two Doublet Higgs Model at the Hadrons Colliders", N.Baket et al.
https://www.researchgate.net/publication/280329743_Neural_Networks_Search_for_Charged_Higgs_Boson_of_Two_Doublet_Higgs_Model_at_the_Hadrons_Colliders. *Last visited:* 07/09-2016

[5] CMS Detector Description & illustration ,
http://cms.web.cern.ch/news/cms-detector-design. *Last visited:* 07/09-2016

[6] CMS Tracking Description http://cms.desy.de/e53612/e155175/e155179/. *Last visited:* 07/09-2016

[7] Larmor-radius, https://en.wikipedia.org/wiki/Gyroradius. *Last visited:* 07/09-2016

[8] "Cluster properties from two-particle angular correlations in p+p collisions at $\sqrt{S_{NN}} = 200$ and 410 GeV", B.Alver, et al, the PHOBOS Collaboration https://arxiv.org/abs/0704.0966. *Last visited:* 07/09-2016

[9] CMS collaboration (2014). "MinimumBias primary dataset in AOD format from RunB of 2010 (/MinimumBias/Run2010B-Apr21ReReco-v1/AOD)". CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.6BPY.XFRQ/ http://opendata.cern.ch/record/8. *Last visited:* 07/09-2016

[10] CMS Draft Note AN-10-012. "Observation of two-particle correlations over a large pseudorapidity gap in high multiplicity proton-proton collisions at $\sqrt{S_{NN}} = 7$TeV", Head Id: 16948, Archive ID 1998:17011MP, Archive Date 2010/09/13.