Measurement of Higgs boson properties in the four-lepton decay channel with the CMS Experiment data at $\sqrt{s} = 13$TeV

Università degli Studi di Torino

Dipartimento di Fisica

Corso di Laurea Magistrale in Fisica Nucleare e Subnucleare

Tesi di Laurea Magistrale

Candidata: Alessandra Cappati

Relatore: Prof. Nicola Amapane

Controrelatrice: Prof.ssa Raffaella Bonino

23 Ottobre 2017

Anno Accademico 2016 - 2017
Abstract

This thesis reports measurements of Higgs boson properties performed in the four-lepton final state ($H \rightarrow ZZ \rightarrow 4\ell$) using proton-proton collision data collected by the CMS experiment at CERN during 2016 and 2017 at a centre-of-mass energy of 13 TeV. This “golden channel” has been widely exploited for the study of the Higgs boson, since the decay kinematics can be fully reconstructed.

My goal for this analysis is to investigate the secondary Higgs boson production modes (Vector Boson Fusion, associated production with a vector boson or with a top-antitop quark pair) by classifying events into mutually exclusive categories according to the corresponding additional signatures. Probing these processes is interesting per se, but also allows couplings of the Higgs bosons to fermions and gauge bosons to be measured in order to test possible deviations from the Standard Model predictions.
## 3.8.1 Trigger requirements

- Content

## 3.8.2 Candidate selection

- Content

## 3.8.3 Choice of the best ZZ candidate

- Content

## 3.8.4 Event selection efficiency

- Content

## 4 The statistical apparatus

### 4.1 A statistical model

- Content
  - 4.1.1 The likelihood function
  - 4.1.2 Incorporating systematic uncertainties

### 4.2 Hypothesis testing

- Content

### 4.3 Quantifying an excess

- Content

### 4.4 Measuring parameters

- Content

### 4.5 Inputs for the statistical model

- Content

## 5 Measurement of Higgs boson properties with 2016 data

### 5.1 Data and simulated samples

- Content

### 5.2 Observables and kinematic discriminants

- Content
  - 5.2.1 Decay observables
  - 5.2.2 Production observables

### 5.3 Categorizing events

- Content

### 5.4 Inputs for the statistical model

- Content
  - 5.4.1 Event yields
  - 5.4.2 Mass lineshape
  - 5.4.3 Systematic uncertainties

### 5.5 Results of $H \to ZZ \to 4\ell$ 2016 analysis

- Content
  - 5.5.1 Event yields
  - 5.5.2 Event distributions
  - 5.5.3 Significance and signal strength modifiers

### 5.6 Study on an alternative event categorization

- Content

## 6 A first look at 2017 data

### 6.1 Data and MC samples

- Content
  - 6.1.1 2017 data sample
  - 6.1.2 2016 data and MC samples

### 6.2 Event selection

- Content

### 6.3 Corrections applied

- Content

### 6.4 Data - MC comparison

- Content

### 6.5 Time dependency study

- Content
  - 6.5.1 Z boson properties
  - 6.5.2 Z boson reconstruction rate
  - 6.5.3 Lepton properties

### 6.6 Effect of FSR recovery

- Content

### 6.7 ZZ mass distributions

- Content
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusions</td>
<td>102</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>104</td>
</tr>
<tr>
<td>Bibliography</td>
<td>105</td>
</tr>
</tbody>
</table>
Introduction

The Higgs boson was introduced in the context of the electroweak symmetry breaking mechanism. This is a mechanism introduced in 1964 by Brout, Englert and Higgs to explain how particles acquire a mass. In fact the Standard Model of particle physics (SM) predicts all particles to be massless, but we observe them with a mass different from zero. This mechanism is built on the Goldstone theorem and on a doublet of complex scalar fields, and describes how particle masses are generated through the interaction with the Higgs field. The main experimental implication is that it predicts the existence of a new physical scalar boson: the Higgs boson.

Since its postulation in the Standard Model, the Higgs boson has been deeply studied both from the theoretical side in order to understand its nature and searched for experimentally in order to observe it.

Various high energy experiments during the years tried to detect the predicted particle, scanning different mass intervals, since the Higgs boson mass is a free parameter of the model. After 48 years, the predicted boson was discovered in July 2012 at the European Centre for Nuclear Research (CERN), with mass near 125 GeV/c$^2$ and properties consistent with the theorized standard model Higgs boson. The observation was reported by two of the CERN Collaborations: ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid).

The discovery relied on a combination of studies in five different decay channels, but two of them provided the most sensitive measurement, thanks to the possibility of fully reconstructing the kinematics of the decay. These two channels are the decay into a pair of photons ($H \rightarrow \gamma \gamma$) and the decay into a pair of $Z$ bosons that both decay into pairs of electrons or muons ($H \rightarrow ZZ \rightarrow 4\ell, \ell = e, \mu$).

After the discovery, further studies have been performed on this new particle in order to measure its properties and to verify the agreement with the Standard Model predictions. These studies were performed in the Run I of the Large Hadron Collider (LHC) at a centre of mass energy of 7 and 8 TeV, and they are being performed in the Run II at $\sqrt{s} = 13$ TeV. The larger production cross-section at higher energy allows properties of the Higgs boson that were not accessible before to be measured. For example secondary Higgs boson production mechanisms, like Vector Boson Fusion (VBF) and associated production with a vector boson (WH and ZH) or with a top-antitop quark pair (t\bar{t}H), can now be investigated.

The CMS Experiment has maximally exploited the four-lepton channel taking advantage
of its large signal-to-background ratio and of the complete reconstruction of the final state to measure several properties of the Higgs particle such as the mass, spin-parity, width, lifetime.

This thesis work is focused on the analysis of the \( H \rightarrow ZZ \rightarrow 4\ell \) process with the LHC Run II data, in the context of the broader analysis effort of the CMS Collaboration on this channel. I initially contributed to the first paper \(^1\) reporting measurements of Higgs boson properties performed with the proton-proton data collected in 2016 by the CMS experiment in the four-lepton decay channel. For this paper I mainly contributed performing the statistical analysis on the collected data in order to extract the results. After the analysis of 2016 data I focused on preparing the analysis of the first 2017 data, since in May 2017 the LHC started again delivering proton-proton collisions. I initially performed the first checks on the quality of the data collected in the first months of 2017 by the CMS detector. The purpose of these first studies is to check possible problems on the efficiency of trigger, object reconstruction or momentum scale, that can happen since the CMS pixel detector has been replaced at the beginning of 2017. For this reason, I considered the distributions of the reconstructed quantities relevant to the analysis and their stability in time, comparing the results with Monte Carlo (MC) simulations and with 2016 data. The goal of these checks is to prepare the environment for the analysis that will be performed with the full 2017 dataset. After these checks, I performed a first analysis on early 2017 data corresponding to an integrated luminosity of 13.88 fb\(^{-1}\), observing the Higgs boson peak in the four-lepton invariant mass distribution and testing the agreement between expected and observed mass distribution.

This thesis is structured as follows. Chapter \(^1\) briefly introduces the theory of the Higgs mechanism and the Higgs boson phenomenology at hadron colliders. Chapter \(^2\) presents an introduction to the experimental apparatus, the LHC accelerator and the CMS detector. Chapter \(^3\) focuses on the golden channel \( H \rightarrow ZZ \rightarrow 4\ell \), presenting its features and the requirements for the event selection. The following Chapter \(^4\) presents the statistical model used to extract properties from the selected data. Chapter \(^5\) then, reports the results of the analysis based on the data sample recorded in 2016. In the end, Chapter \(^6\) reports the first results obtained with the 2017 data.
Chapter 1

The Higgs boson in the Standard Model

1.1 The Standard Model

The Standard Model of particle physics (SM) is a quantum field theory which describes the fundamental constituents of matter and their fundamental interactions. In particular, this model describes the electroweak and strong interactions between fundamental particles and it is the theory that nowadays best explains the experimental observations of sub-nuclear phenomena.

In the Standard Model, matter is described by twelve fundamental spin-$\frac{1}{2}$ fermion fields: six for quarks, which are subject to both strong and electroweak interactions and do not exist as free states, and six for leptons, which are sensitive to the weak and electromagnetic forces. These fermion fields are organized in three generations: the first one, that comprehends up and down quarks and the electron, constitutes ordinary matter.

The Standard Model describes the strong and the electroweak interactions between fundamental particles as the result from the exchange of force-carrier particles: the vector bosons. Each fundamental force has its own corresponding bosons: photon and Z, $W^+$, $W^-$ bosons for the electroweak interaction, gluons for the strong force. Mathematically, the fundamental forces are described with the groups algebra, and the Standard Model describes the electroweak and strong interactions using the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$, where:

- $SU(3)_C$ is the colour group of Quantum Chromodynamics (QCD), which is the theory that describes the strong interaction, involving eight gluon gauge fields;

- $SU(2)_L \times U(1)_Y$ is the weak isospin and hypercharge group of the electroweak theory, which unifies electromagnetism and weak interaction, and involves four gauge fields $W^{1,2,3}_\mu$ and $B_\mu$. These gauge fields can be transformed into the physical fields $W^\pm_\mu$, $Z_\mu$, that correspond to the three weak bosons, and $A_\mu$ that corresponds to the photon.
Figure 1.1 shows the fundamental particles of the Standard Model in its current formulation: the three generations of fermions which constitute matter, the gauge bosons which mediate the fundamental interactions and a scalar boson which is the focus of this thesis.

1.2 The mass issue

The scalar sector we have just introduced in the previous section is needed to explain the origin of the masses of fermions and gauge bosons. All the fundamental particles of the Standard Model, in fact, have a mass different from zero. But the simple addition of the mass term in the Lagrangian violates the gauge invariance of the theory. For example a mass term for a W boson can be written like $m_W^2 W^\mu W^\mu$, and it is not invariant under gauge transformations of the $SU(3)_C \times SU(2)_L \times U(1)_Y$ group. In fact, a gauge transformation on this term introduces additional terms in the Lagrangian, which spoil the gauge invariance of the Lagrangian itself. Furthermore, fermions in the Standard Model are chiral and the left-handed and right-handed fermions are described by different representations of the $SU(2)$ group. Therefore also mass terms of fermions are not invariant under gauge transformations. Then the
electroweak theory predicts the weak gauge bosons and the fermions to be massless, because adding \textit{ad hoc} mass terms would violate the gauge invariance of the Lagrangian. Particles masses, therefore, have to be provided by a different mechanism. The solution to the mass problem was proposed in 1964 simultaneously in three independent papers from Higgs \cite{2}, Englert and Brout \cite{3}, and Guralnik, Hagen and Kibble \cite{4}. The mechanism introduced is known today as Brout-Englert-Higgs (BEH) mechanism. It provides the masses of the Standard Model particles through the interaction of SM particles with a scalar field, the so called Brout-Englert-Higgs (BEH) field.

The idea of the BEH mechanism is that the entire universe is filled with an homogeneous field, the BEH field, and particles are in principle massless. Their interaction with the field provides them a mass and each particle obtains a mass proportional to the magnitude of its interaction with the BEH field\footnote{In the case of neutrinos, it is possible that the electroweak symmetry breaking mechanism plays only a partial role in generating the observed neutrino masses, with additional contributions at a higher scale via the so called see-saw mechanism.}

Having in mind the Englert, Brout and Higgs idea, the following section presents this mechanism with a more formal approach\cite{5}.

\section{The Brout-Englert-Higgs mechanism}

The BEH mechanism postulates the existence of a new field which is symmetric under the gauge transformations, and acquires a non-zero expectation value in the vacuum state, breaking the electroweak symmetry. The simplest field that we can introduce is a $SU(2)_L$ doublet of complex scalar fields:

\begin{equation}
\phi = \begin{pmatrix} \phi^+ \\ \phi^- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}
\end{equation}

which is introduced in the Lagrangian density of the SM via the term

\begin{equation}
L_{BEH} = (D^\mu \phi)^\dagger (D_\mu \phi) + V (\phi^\dagger \phi)
\end{equation}

where $D_\mu = \partial_\mu + \frac{i}{2} g \sigma_a W^a_\mu + \frac{i}{2} g' Y B_\mu$ is the covariant derivative.

The potential is chosen as (see Figure 1.2):

\begin{equation}
V (\phi^\dagger \phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2
\end{equation}

where $\lambda > 0$ is required for the vacuum to be stable, while $\mu^2 < 0$ is chosen in order to induce spontaneous symmetry breaking.

The ground state is given by:

\begin{equation}
\phi^\dagger \phi = -\frac{\mu^2}{2\lambda}
\end{equation}
This condition determines only the length of the vector $\phi$, its direction is arbitrary. Therefore, if we choose to fix this vacuum state on the $\phi^-$ axis and we expand the $\phi$ field around the minimum of the potential, the first term of the expansion will be:

$$
\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, v^2 = -\frac{\mu^2}{\lambda} \tag{1.5}
$$

where $v$ is called the vacuum expectation value of $\phi$. It is possible to prove that this state is invariant under the $U(1)_{\text{EM}}$ symmetry. The field can be re-parametrized as:

$$
\phi(x) = \frac{1}{\sqrt{2}} e^{i\theta^a(x)\sigma_a} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, a = 1, 2, 3 \tag{1.6}
$$

where $\theta^1, \theta^2, \theta^3, h$ are four real scalar fields, the apparition of which is predicted by the Goldstone theorem. The $\theta^a$ fields can be eliminated by choosing an appropriate gauge: they become the longitudinal degrees of freedom of the $W^\pm$ and $Z^0$ vector bosons, that initially were massless and therefore had only two transversal degrees of freedom. This leaves the vacuum state containing a single scalar field $h$, the quanta of which are associated to the physical particle Higgs boson.

Inserting into Eq. 1.2 the expression of $\phi(x)$ and of the covariant derivative provides the mass terms for the weak vector bosons, while the photon field remains massless. The Higgs boson also acquires a mass:

$$
m_H = \sqrt{2} |\mu| \tag{1.7}
$$

which is a free parameter of the theory. Additional trilinear and quadrilinear terms in the Lagrangian describe the self-coupling of the Higgs boson and its couplings to weak vector bosons.

The spontaneous breaking of the $SU(2)_L \times U(1)_Y$ symmetry into the $U(1)_{\text{EM}}$ one also allows to extend the SM Lagrangian with gauge-invariant terms that generate masses for the fermions. These so-called Yukawa terms take, for example, the form $\bar{\psi}_L \psi_R \phi$.
down-type fermions. After electroweak symmetry breaking, the additional terms in the Lagrangian take the form:

\[ \mathcal{L}_{Yukawa} = - \sum f m_f \bar{\psi} \psi \left( 1 + \frac{h}{v} \right) \]  \hspace{1cm} (1.8)

where \( m_f \) are mass matrices that account for the fact that the weak eigenstates and mass eigenstates of fermions differ. This procedure implies that the Higgs field interacts with fermions, with couplings proportional to fermion masses.

The BEH mechanism, proposed in 1964, received an experimental confirmation through the discovery in 2012 at the CERN Large Hadron Collider (LHC) of a particle consistent with the Higgs boson that it predicts.

1.4 Higgs boson phenomenology at a proton-proton collider

1.4.1 Higgs boson production mechanisms

The Higgs boson can be produced with different mechanisms. At the Large Hadron Collider (LHC) the final states of these production modes can involve other particles than just the Higgs boson, as shown in Figure 1.3. This leads to different experimental signatures which help identifying these processes.

- **Gluon Fusion** (ggH): this process involves two gluons that merge into a Higgs boson via an intermediate quark loop (Figure 1.3 top left). Quarks of all flavours contribute to the loop, but contributions from lighter quarks are suppressed proportionally to \( m_q^2 \), therefore the top quark, which is the most massive one, gives the largest contribution to the loop. ggH is the dominant Higgs boson production mechanism at the high centre-of-mass energies provided by LHC, because the gluon density is very large at these energies. This mechanism dominates over all others by more than one order of magnitude, as shown in Figure 1.4.

- **Vector Boson Fusion** (VBF): this is the production mode with the second-largest cross section at the LHC. Its cross section is 12 times smaller than the one for ggH. At the leading order, it consists in the scattering of two (anti-)quarks, mediated by the exchange of a vector boson (\( W^\pm \) or \( Z^0 \)) which radiates off the Higgs boson (Figure 1.3 top right). The two scattered quarks hadronize in hard jets in the forward and backward regions of the detector. These jets constitute a very clean experimental signature.

- **Associated production with a vector boson** (VH): often referred as Higgstrahlung, is the third most relevant process at LHC. It can be split in two processes, ZH and WH, according to the flavour of the associated vector boson. In this process a quark and an antiquark merge into a vector boson (\( W^\pm \) or \( Z^0 \)) that
Chapter 1. The Higgs boson in the Standard Model

Figure 1.3: Leading order Feynman diagrams for the four dominant production modes of the SM Higgs boson at LHC: gluon fusion (top left), vector boson fusion (top right), VH associated production (bottom left) and t\bar{t}H associated production (bottom right).

radiates the Higgs Boson (Figure 1.3 bottom left). The VH experimental signature is characterized by the Higgs boson decay products accompanied by the products of the associated vector boson. When this vector boson decays hadronically, it provides a pair of nearby boosted jets with invariant mass close to the nominal mass of $W^\pm$ or $Z^0$; when it decays in leptonic ways, it provides one lepton and missing transverse energy (for WH), or a pair of leptons or missing transverse energy (for ZH).

- **Associated production with a top quark pair** (t\bar{t}H): this process is gluon induced and in the final state the Higgs boson is accompanied by a t\bar{t} pair (Figure 1.3 bottom right). Each associated top quark then decays into a bottom quark and a $W^\pm$ boson which can decay leptonically or hadronically, leading to several possible experimental signatures. The t\bar{t}H process is about 100 times rarer than ggH, but it allows the Yukawa coupling between the top quark and the Higgs boson to be proven directly.

- **Associated production with a bottom quark pair** (b\bar{b}H): this is gluon induced and in the final state the Higgs boson is accompanied by a b\bar{b} pair. It presents a cross section of a similar order of magnitude and a similar Feynman
1.4.2 Higgs boson decay modes

The Higgs boson directly couples to all massive particles of the Standard Model and can also couple to massless particles via intermediate loops. This fact leads to a variety of different decay channels.

The total decay width of the Higgs boson and the relative branching fractions of its decay channels are determined by the value of its mass. Figure 1.5 (left) presents the values of the branching fractions as a function of the hypothesized Higgs boson mass \( m_H \). A Higgs mass of about 125 GeV implies an interesting succession of decay channels, as shown in Figure 1.5 (right). In particular the dominant decay modes are \( H \to \bar{b}b \) and \( H \to WW^* \), followed by \( H \to gg \), \( H \to \tau^+\tau^- \), \( H \to c\bar{c} \) and \( H \to ZZ^* \). The Higgs decay channels \( H \to \gamma\gamma \), \( H \to \gamma Z \) and \( H \to \mu^+\mu^- \) follow with much smaller branching ratios. All these decay modes provide an excellent opportunity to explore the Higgs couplings to many SM particles. In particular channels like:

- \( H \to \bar{b}b \), \( H \to \tau^+\tau^- \), \( H \to c\bar{c} \) and \( H \to \mu^+\mu^- \) allow Higgs boson direct couplings to fermions to be investigated;
Chapter 1. The Higgs boson in the Standard Model

Figure 1.5: SM Higgs boson branching ratios and their uncertainties as a function of $m_H$, in a low-mass range (left) [9] and for the mass range around 125 GeV (right) [8].

- $H \rightarrow WW^*$ and $H \rightarrow ZZ^*$ give information on Higgs boson couplings to weak vector bosons.

For the Higgs boson discovery, five of all the decay channels played an important role at the LHC, and they are still studied nowadays. Their importance is due both to the magnitude of their branching fraction and on the experimental possibility of efficiently extracting the corresponding signals while rejecting their backgrounds. The $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4\ell$ channels gained the definition of **golden channels**, thanks to the excellent Higgs boson mass resolution and to the possibility of fully reconstructing the kinematics of the event. The $H \rightarrow W^+W^- \rightarrow \ell^+\nu\ell^-\bar{\nu}$ channel, instead, beside the relatively large branching ratio, has a poor mass resolution due to the presence of neutrinos. The $H \rightarrow b\bar{b}$ and $H \rightarrow \tau^+\tau^−$ channels, in the end, suffer from large backgrounds and a poor mass resolution.

1.5 The Higgs boson discovery

The Higgs boson is the main experimental implication of the BEH mechanism, and its experimental search has been carried on for more than 40 years. The Higgs boson mass $m_H$ is a free parameter of the model, therefore various searches scanned different mass intervals. The allowed range for the Higgs boson mass was restricted from theoretical arguments and got progressively constrained by experimental searches. In fact, the existence of the Higgs boson was excluded in large intervals of mass by the CERN Large Electron Positron collider (LEP) and the Fermilab Tevatron in the 1990s and early 2000s. Afterwards the existence of the Higgs boson was investigated at the CERN Large Hadron Collider (LHC), which was intended to provide proton-proton (pp) collisions with a
nominal centre-of-mass energy of 14 TeV. The LHC began delivering pp collisions in 2009, reached conditions suitable for physics in 2010, and finally delivered two high-luminosity data samples in 2011 and 2012, at centre-of-mass energies of 7 TeV and 8 TeV, respectively. This initial data taking period is referred to as Run I. After the Higgs boson existence was excluded in new regions of the mass range in late 2011, on the 4th July 2012 the discovery of a new boson with mass around 125 GeV/c² and properties consistent with the predicted Standard Model Higgs boson was announced by the ATLAS [10] and CMS [11] Collaborations.

The discovery relied on a combination of studies in different decay channels, but the two of them which provided most of the sensitivity and a measurement of the mass are the decay to a pair of photons (H → γγ) and the decay to a pair of Z bosons that then decay into pairs of electrons or muons (H → ZZ → 4ℓ, ℓ = e, µ). Figure 1.6 shows the invariant mass distribution obtained, by the CMS Collaboration, for these two final states. In these two plots the Higgs boson is shown as an excess around 125 GeV/c².

With the additional data collected in 2012, the new boson properties started to be investigated. It was found to be consistent with the theoretical prediction for the spin-parity J^P = 0^+ and at the end of the LHC Run I its mass was measured to be m_H = 125.09 ± 0.21 (stat.) ± 0.11 (syst.) GeV/c² [13], with a combination of the results obtained by ATLAS with 4.9 fb⁻¹ at 7 TeV and 20.2 fb⁻¹ at 8 TeV and CMS with 5.1 fb⁻¹ at 7 TeV and 19.7 fb⁻¹ at 8 TeV. Other Higgs boson properties were also studied as cross-section, couplings with fermions and bosons and width [14],[15],[16].
1.6 Perspectives

In 2015 the LHC started again to deliver proton-proton collisions, but at higher centre-of-mass energy ($\sqrt{s} = 13$ TeV) starting the data taking period referred to as Run II. The purpose of this new run is to better measure the Higgs boson properties, focusing in particular on less accessible characteristics. Among these there are the sub-dominant production modes of the Higgs boson: vector boson fusion and associated production with a vector boson or a top-antitop quark pair.

Studying Higgs boson secondary production mechanisms allows some Higgs boson couplings to be measured, like couplings with vector bosons or the Yukawa coupling with the top quark. These new studies on the Higgs boson are necessary to further understand its nature, to achieve a full knowledge of its properties, and to test possible deviation from the Standard Model predictions.

The present thesis focuses on the analysis of data collected by the CMS detector in the golden channel $H \rightarrow ZZ^* \rightarrow 4\ell$. In particular the analysis of 2016 data corresponding to 35.9 fb$^{-1}$ is presented in Chapter 5 investigating Higgs boson secondary production modes. These measurements are performed exploiting the large integrated luminosity available, which allows less accessible properties to be studied. The higher the integrated luminosity the bigger the collected amount of data, and the expected integrated luminosity at the end of LHC Run II ($\sim 150$ fb$^{-1}$) will allow all the rare properties of the Higgs boson to be reached. In this perspective, Chapter 6 presents the first controls made on the first 2017 data (corresponding to 13.88 fb$^{-1}$), that are aimed at probing the quality of collected data and at testing the environment built for data analysis which will be performed with the full 2017 dataset.

Since the analysis is performed in the context of the CMS Experiment, the following Chapter 2 will briefly illustrate the CMS detector.
Chapter 2

The experimental set-up

This thesis is realized within the CMS Experiment, exploiting data from proton-proton (pp) collisions delivered by the CERN Large Hadron Collider (LHC).

The CERN, the European Organization for Nuclear Research, is a scientific laboratory founded in 1954 by 12 European countries to investigate the physics of atomic nuclei. Soon the understanding of matter went deeper and deeper than the nucleus, and today the CERN’s main area of research is particle physics: the study of the fundamental constituents of matter and the forces acting between them.

In order to study the fundamental laws of nature, particles are made to collide together at close to the speed of light velocity. The instruments used for accelerating particles to this speed are particle accelerators. They boost beams of particles to high energies before the beams are made to collide with each other. Then, the results of these collisions are observed and recorded by particle detectors. To reach close to the speed of light velocity, particles go through a long chain of acceleration that, today, ends with the Large Hadron Collider (LHC).

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) \[17\] is the world’s largest and most powerful particle accelerator. It consists in a 27-kilometre two-ring, superconducting accelerator and collider, designed to collide proton beams up to a nominal centre-of-mass energy of \(\sqrt{s} = 14\) TeV (i.e. 7 TeV per beam) and an instantaneous luminosity of \(10^{34}\) cm\(^2\)s\(^{-1}\). The beams travel in opposite directions in separate beam pipes which are two tubes kept at ultrahigh vacuum. Beams are guided around the accelerator ring by a strong magnetic field maintained by superconducting electromagnets made by coils of copper-clad niobium-titanium. Magnets operate in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires superfluid helium in order to cool the magnets and maintain them at the operating temperature of 1.9 K (\(-271.25^\circ\)C).

Thousands of magnets of different varieties and sizes are used for different purposes: dipole magnets are used to bend the beams around the accelerator, quadrupole magnets
Chapter 2. The experimental set-up

Figure 2.1: The CERN accelerator complex [18]. The proton injection chain for the LHC starts from the LINAC2 and proceeds through the Booster, PS, and SPS.

focus the beams and sextupole magnets are used to squeeze the beams further close to the intersection points to maximize the probability of interactions.

Before being injected into the LHC the proton beams are prepared by a chain of pre-accelerators which increase their energy in steps. This system is presented in the Figure 2.1 which shows the CERN accelerator complex. Protons are first accelerated to an energy of 50 MeV in the Linear Accelerator (LINAC2), which feeds the Proton Synchrotron Booster (PSB), where protons are accelerated up to 1.4 GeV. Then particles reach 26 GeV in the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS) further increases their energy to 450 GeV. The protons finally are injected in the two beam pipes of the LHC, where in one of them the beam circulates clockwise while in the other one the beam circulates anticlockwise. It takes about four minutes to fill each LHC ring and about 20 minutes for the protons to reach their maximum energy of 6.5 TeV. Beams circulate for many hours inside the LHC beam pipes under normal operating conditions.
Chapter 2. The experimental set-up

The two beam pipes are brought into collision inside four detectors (ALICE, ATLAS, CMS and LHCb) where the total energy at the collision point is 13 TeV. Protons are not the only particles accelerated in the LHC. Lead ions for the LHC are taken from a source of vaporized lead and then enter in Linac 3 before being collected and accelerated in the Low Energy Ion Ring (LEIR). They then follow the same journey to maximum energy as the protons.

Protons circulate in the LHC in bunches spaced by 25 ns (or 7.5 m), therefore the bunch crossing rate is 40 MHz. The nominal value of protons per bunch is \( N_b = 12 \times 10^{11} \), and the nominal value of bunches per beam is \( n_b = 2808 \). Under nominal conditions, the number of inelastic collision events is of the order of \( 10^9 \) per second, with \( \sim 20 \) collisions per bunch crossing.

The LHC instantaneous luminosity depends on the beam parameters and can be written as:

\[
L = \frac{f_{\text{rev}} N_b^2 n_b \gamma_r F}{4\pi \epsilon_n \beta^*}
\]

where \( f_{\text{rev}} = 11 \text{ kHz} \) is the revolution frequency, \( \epsilon_n \) the normalized transverse beam emittance, \( \beta^* = 0.55 \text{ m} \) is the nominal value of the beta function at the collision point, which measures the beam focalization and is corrected by the relativistic gamma factor \( \gamma_r \), and \( F \) is a geometric luminosity reduction factor that accounts for the crossing angle at the interaction point.

The nominal instantaneous luminosity of LHC is \( 10^{34} \text{ cm}^2\text{s}^{-1} \) and thanks to this the integrated luminosity delivered by LHC in 2016 and up to mid-September 2017 amount to 35.9 fb\(^{-1}\) and 13.88 fb\(^{-1}\) respectively. These are the sample considered for the analysis presented in this thesis in Chapter 5 and Chapter 6 respectively.

Four main particle detectors are installed in underground caverns at the four beam intersection points. The two largest ones, ATLAS (A Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) are designed to cover a wide physics program in the scalar, electroweak, and strong sectors, with optimized sensitivity for Higgs bison searches and additional possible new physics at the TeV scale. The two other detectors are LHCb (LHC beauty), which is aimed at studying CP violation in B-hadrons, and ALICE (A Large Ion Collider Experiment), which is dedicated to heavy-ion collisions in order to study the quark-gluon-plasma.

This thesis is performed in the context of the CMS experiment which is presented in the following section.

2.2 The Compact Muon Solenoid

The CMS detector layout is organized around a superconducting solenoid magnet of 6 m internal diameter and 12.5 m length, which provides a large magnetic field of 3.8 T. The overall apparatus is rather compact: it materializes as a 21.6-metre-long, 14.6-metre-wide, 14000-tonne cylinder around the LHC beam axis. Within the solenoid volume
Chapter 2. The experimental set-up

there are three major subsystems: a silicon pixel and strip tracker which measures the trajectories of charged particles, a lead tungstate crystal electromagnetic calorimeter (ECAL) that mainly collects the energies of electrons and photons, and a brass and scintillator hadronic calorimeter (HCAL) which stops the more penetrating hadrons. Some forward calorimeters further improve hermeticity. The measurement of muons relies on a combination of inner tracking and information from the muon chambers, which are gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid. The overall layout of the detector is shown in Figure 2.2. A detailed description of the CMS detector can be found in Ref [19].

2.2.1 The coordinate system

A conventional coordinate system has been adopted to describe the CMS detector, and will be used throughout this thesis. The origin of the coordinate system is located at the nominal collision point. The $z$ axis coincides with the proton beam direction and points toward the Jura mountains from the LHC Point 5 where the CMS experiment is located. The $y$ axis points vertically upward, while the $x$ axis points radially inward toward the centre of the LHC ring. The azimuthal angle $\phi$ is defined from the $x$ axis in the $xy$ plane and it takes values between $-\pi$ and $\pi$. The radial coordinate in the $xy$ plane

Figure 2.2: A perspective view of the CMS detector, illustrating its major subsystems.
is labelled \( r \). The polar angle \( \theta \) is defined from the \( z \) axis and it assumes values from 0 to \( \pi \). Usually the polar coordinate is quoted using another variable, the pseudorapidity, which is defined as \( \eta \equiv -\ln[\tan(\theta/2)] \).

From these coordinates it is possible to define some quantities which will be used in the analysis. Two of these quantities of interest are the particle momentum and energy transverse \( p_T \) and \( E_T \) which are computed using the components of the momentum and energy in the \( xy \) plane as:

\[
p_T = \sqrt{p_x^2 + p_y^2} \tag{2.2}
\]
\[
E_T = \sqrt{m^2 + p_T^2}
\]

where \( m \) is the mass of the particle and \( p_x \) and \( p_y \) are the \( x \) and \( y \) components of the particle momentum, respectively.

Another quantity which will be used in the analysis is the missing energy transverse. It is defined as the imbalance of the total transverse energy measurement in a collision and it is labelled as \( E_T^{\text{miss}} \).

Finally, the angular distance between two particles \( i \) and \( j \) is defined as:

\[
\Delta R(i, j) = \sqrt{(\eta^i - \eta^j)^2 + (\phi^i + \phi^j)^2} \tag{2.3}
\]

Based on the \( \eta \) coordinate, the detector is divided into a central part called the barrel and two opposite forward parts called endcaps. The exact boundaries between these regions depend on the subsystems.

### 2.2.2 The tracking system

The tracking system is the most inner detector of the CMS Experiment. It surrounds the interaction point, has a length of 5.8 m, a diameter of 2.5 m and is fully immersed in the magnetic field of 3.8 T provided by the CMS solenoid. It is designed to provide a precise and efficient measurement of the trajectories of charged particles emerging from the LHC collisions and a precise reconstruction of secondary vertices.

Since several hundreds of particles go through the tracker during each bunch crossing, high granularity and fast response are especially important to its design, in order to correctly identify particle trajectories and reliably attribute them to the correct bunch crossing. All this implies a high power density of on-detector electronics, which in turn requires efficient cooling. Nevertheless, the amount of material has to be kept to the minimum, in order to limit phenomena such as multiple scattering, bremsstrahlung and photon conversion which complicate particle reconstruction. The request of radiation resistance is also necessary, since the tracker surrounds the interaction point.

In order to satisfy all these requests the tracking system has been built with silicon detectors. The overall layout of this system is shown in Figure 2.3.
Figure 2.3: Schematic longitudinal section of the CMS tracker, showing inner pixel detector with its barrel and endcap modules, and the strip detector with two collections of barrel modules, the tracker inner barrel (TIB) and the tracker outer barrel (TOB), and two collections of endcap modules, the tracker inner discs (TID) and the tracker endcaps (TEC) [19].

The CMS tracker is based on two types of sensors: silicon pixel sensors of size $100 \times 150 \mu m^2$ displayed on the three cylindrical barrel layers, and on two endcap disk layers, and silicon strip sensors displaced in the rest of the tracker, arranged in 10 barrel layers and 12 endcap layers. The tracker detector covers a pseudorapidity range of $|\eta| < 2.5$ allowing the measurement of all the charged particles within this range.

2.2.3 The electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) is designed to provide the measurement of electrons and photons energies. It is a hermetic homogeneous calorimeter made of scintillating crystals of lead tungstate (PbWO$_4$). Thanks to the short radiation length (0.89 cm) and the high density of this material, electromagnetic showers can be absorbed within relatively short crystals, while the small Molière radius (2.2 cm) allows a good shower separation to be obtained. The 80% of the scintillation light is emitted in 25 ns and this allows a detector response fast enough to cope with the LHC bunch spacing. The ECAL general structure is shown in Figure 2.4.

The barrel part of the ECAL (EB) covers the pseudorapidity range $|\eta| < 1.479$ and it is made of crystal of length 230 mm and and frontal cross section $22 \times 22 mm^2$ organized in 36 supermodules which cover half of the barrel length and 20° in $\phi$. Each supermodule is made of four modules which contain 400 or 500 crystals in an alveolar structure.

The ECAL endcaps (EE), instead, covers the pseudorapidity range $1.479 < |\eta| < 3.0$ and it is made of crystals with length 220 cm and frontal cross section $28.62.62 mm^2$.
Chapter 2. The experimental set-up

Figure 2.4: Layout of the CMS electromagnetic calorimeter showing the arrangement of crystal modules, supermodules and endcaps, with the preshower in front [19].

organized in two semi-circular dees containing 3662 crystals.
The crystals are mounted in a quasi-projective geometry, so that their axes make a 3° angle with respect to the direction of the nominal interaction point in both the \( \eta \) and \( \phi \) projections, thus avoiding the alignment of inter-crystal gaps with particle trajectories. The scintillation light from ECAL crystals is read by fast, radiation-tolerant photode-
tectors which can operate inside the CMS magnetic field and are insensitive to particles traversing them.

Two preshower detectors are installed at each end of the tracker detector, in front of the ECAL endcaps, in order to help distinguishing \( \pi^0 \rightarrow \gamma\gamma \) decays from single photons and identifying electrons against minimum ionizing particles. These sampling calorimeters are made of a lead radiator layer which initiates electromagnetic showers from incoming particles, followed by silicon strip sensors that measure the deposited energy.

Although lead tungstate crystals are radiation resistant, they undergo a limited but rapid loss of optical transmission under irradiation. This phenomenon depends on the luminosity and crystal pseudorapidity, and is partly balanced by an annealing effect. This effect is measured and it is taken into account by time-dependent corrections applied to the measured particle energies.
Chapter 2. The experimental set-up

2.2.4 The hadronic calorimeter

The hadron calorimeter (HCAL) is designed to measure the energy of hadrons, therefore it is important for the reconstruction of particle jets. Its wide extension in pseudorapidity captures a large fraction of particles coming from the interaction point and this allows a reliable measurement of missing transverse energy to be performed which is a signature of otherwise undetected particles such as neutrinos.

The structure of the CMS HCAL is illustrated in Figure 2.5. The HCAL barrel (HB) is located between the outer extent of the EB and the inner extent of the magnet coil. This extension is not enough to fully absorb the hadronic showers, therefore an outer hadron calorimeter (HO) is placed outside of the solenoid. The HCAL barrel covers a pseudorapidity range $|\eta| < 1.3$, which the HCAL endcaps (HE) extend to $|\eta| < 3.0$. In addition, two forward hadron calorimeters (HF) are located 11.2 m away from the interaction point and extend the pseudorapidity coverage up to $|\eta| < 5.2$, thus ensuring good hermeticity.

The HCAL is a sampling calorimeter. In the HB and HE, brass layers are used as the absorber material and are interspersed with plastic scintillator tiles which are the active material. The HB and HE calorimeter cells are grouped in projective towers and scintillation light is converted by wavelength-shifting fibres embedded in the scintillator.

Figure 2.5: Longitudinal section of a quarter of the CMS detector, showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters [19].
tiles and detected by hybrid photodiodes. The HFs instead have to sustain a harsher radiation environment therefore quartz fibres are used as active material. They are placed between steel absorber plates and emit Cherenkov light, which is detected by photomultipliers.

2.2.5 The muon system

Muons play a major role in many physics analyses, and in particular in the study of the golden channel $H \rightarrow ZZ \rightarrow 4\ell$. The CMS muon system has been designed to achieve high-precision measurement of muon momenta and charge. As shown in Figure 2.6 the muons system is made of muon chambers embedded in the iron return yoke of the CMS magnet and it is divided into a cylindrical barrel section and two planar endcap regions. Three types of gas-ionization chambers are used in this system: Drift Tubes, Cathode Strip Chambers and Resistive Plate Chambers.

The Drift Tubes (DTs) are located in the barrel region, where the muon rate is low, and the magnetic field is quite uniform, and cover a pseudorapidity range of $|\eta| < 1.2$. The DTs are organized into four stations interspersed among the layers of the flux return
plates. Their basic constituents are rectangular drift cells bounded by two parallel aluminium planes, which serve as cathodes. Anodes are 50 $\mu$m stainless steel wires located in the centre of the cells. A muon passing through a cell ionises the gas mixture that fills the cell volume. The drift time of the resulting electrons is then used to measure the distance between the muon track and the wire. Each chamber has a resolution of 100 $\mu$m in the $r$-$r - \phi$ plane. The number and orientation of chambers in each station is chosen in order to link muon hits from different stations into a single muon track.

The *Cathode Strip Chambers* (CSCs) are used in the endcaps, where the muon rates and background levels are high and the magnetic field is large and non uniform, and cover a pseudorapidity range of $0.9 < |\eta| < 2.4$. The CSCs are multiwire proportional chambers, made of 6 anode wire planes interleaved among 7 cathode panels, with the wires running approximately perpendicular to the strips. A muon passing through a chamber generates an avalanche, inducing a charge on several cathode strips. The ensuing interpolation allows a spatial resolution of 50 $\mu$m to be obtained. Four stations of CSCs are located in each endcap. The chambers are positioned perpendicular to the beam line and interspersed between the magnetic field flux return plates.

The *Resistive Plate Chambers* (RPCs) are located both in the barrel and in the endcaps and cover the pseudorapidity range $|\eta| < 1.6$. The RPCs are double-gap chambers which operate in avalanche and are disposed in six layers in the barrel and three layers in the endcaps. They provide a complementary trigger system with moderate spatial resolution but excellent time resolution (of the order of 1 ns), which helps measuring the correct beam-crossing time.
Chapter 3

The Golden Channel $H \rightarrow ZZ^* \rightarrow 4\ell$

The decay channel $H \rightarrow ZZ^* \rightarrow 4\ell$ is one of the channels that were most exploited for the Higgs boson discovery and for the measurement of its properties, such as mass, spin-parity and decay width, by the ATLAS and CMS Collaborations at the LHC during the Run I period. In spite of its low branching ratio, the $H \rightarrow ZZ^* \rightarrow 4\ell$ channel presents some features that help the data analysis, for example as:

- the possibility of reconstructing the full final state of the decay, which provides informations on the Higgs boson invariant mass and on decay angles of the decay products;
- the good momentum resolution for leptons (electrons and muons) in the LHC detectors, which allows the Higgs boson resonance to be observed as a narrow peak on the four-lepton invariant mass distribution;
- the large signal-to-background ratio, which allows the signal peak to be measured with a high significance.

Nowadays, in the LHC Run II, the $H \rightarrow ZZ^* \rightarrow 4\ell$ channel is still exploited and, thanks to the increment of the centre-of-mass energy from 7-8 TeV (Run I) to 13 TeV (Run II) and the large amount of integrated luminosity delivered by LHC, some new properties of the Higgs boson become accessible such as its rare production modes. In particular this thesis exploits proton-proton collision data collected by the CMS Experiment in 2016 and 2017 (new data are now being collected) to preform a measurement of properties of the Higgs boson at the larger centre-of-mass energy of $\sqrt{s} = 13$ TeV. Particular emphasis is put on measurements of the Higgs boson couplings with fermions and bosons, investigating its secondary production mechanisms i.e. Vector Boson Fusion, VH associated production and $t\bar{t}H$ production. In this channel the Higgs boson decays to a pair of $Z$ bosons that in turn decay to pairs of leptons. In the current analysis, the four leptons resulting in the final state are only muons or electrons. The $Z$ boson decay into a pair of $\tau$ leptons is not considered because
Chapter 3. The Golden Channel $H \to ZZ^* \to 4\ell$

of the difficulty of identifying this particle due to its hadronic decay modes. Therefore the $H \to ZZ^* \to 4\ell$ channel involves three possible final states: $4e$, where both Z bosons decay into electron-positron pairs, $4\mu$, where both Z bosons decay into $\mu^+\mu^-$ pairs, and $2e2\mu$, where one Z boson decays into $\mu^+\mu^-$ while the other one decays into $e^+e^-$. The background sources to this process include:

- an irreducible four-lepton contribution from the production of $ZZ$ or $Z\gamma^*$ pairs, via the quark-antiquark annihilation ($q\bar{q} \to ZZ$) and the gluon fusion ($gg \to ZZ$) processes. This background is named “irreducible” because it presents the same final state signature as the signal and even an hypothetically perfect detector could not distinguish this background from the signal;

- a reducible contribution from processes which contain non-prompt or misidentified leptons among the four required ones. The main sources of such leptons are non-isolated electrons or muons coming from decays of heavy-flavour mesons, mis-reconstructed jets and electrons from photon conversions. This background is labelled with $Z + X$, where $X$ identifies all the sources of non-prompt or misidentified leptons. $Z + X$ is named “reducible background” because with an hypothetically perfect detector, that could always properly identify leptons, this background could be reduced.

Since the estimation of the reducible background, the resolution, and the systematic uncertainties are different for each of the three possible final state configurations, the three sub-channels $4e$, $4\mu$ and $2e2\mu$ are analysed separately and then the results are combined.

Since the $H \to ZZ^* \to 4\ell$ decay channel presents four leptons in the final state, it is necessary to describe how leptons are reconstructed in the CMS detector and the how events are selected in order to perform the analysis.

3.1 Lepton reconstruction in CMS

All collision events recorded by the CMS detector are centrally processed through a set of reconstruction algorithms that take raw detector data as input, and provide as output a collection of particle candidates with associated properties such as momentum and charge. Event reconstruction in CMS relies on the Particle-Flow (PF) algorithm [21] which is aimed at reconstructing and identifying each stable particle in the event through a combination of all CMS sub-detectors in order to determine their direction, energy and type. The PF candidates are classified as charged hadrons, neutral hadrons, photons, electrons, or muons, and they are then used to build higher-level observables such as jets and lepton isolation quantities. Figure 3.1 illustrates the signatures of the main types of particles passing through the CMS detector.

In the following sections we consider the reconstruction of leptons (electrons and muons), which are fundamental for the analysis of the $H \to ZZ \to 4\ell$ decay channel.
3.1.1 Electron reconstruction

Electrons deposit most of their energy in the ECAL, but since they are charged particles, they also leave hits in the inner tracker. Electron reconstruction in CMS therefore uses an algorithm that combines the inputs from both subsystems. It associates a reconstructed track with a cluster of energy in the ECAL and exploits both sides of the information to estimate the electron momentum.

One of the main difficulties of electron reconstruction in CMS is the amount of tracker material located between the collision point and the ECAL. This causes significant bremsstrahlung along the electron trajectory, with the resulting photons possibly converting to electron pairs. Dedicated techniques have been developed to accurately account for this effect, and will be shown in what follows.

The electron energy usually spreads out over several crystals of the ECAL, and the first step of reconstruction consists in clustering these energy deposits. The algorithm used in Run II for dealing this issue is called PF clustering and consists of three steps. First, cluster seeds are identified as local crystal energy maxima above a given threshold. Second, topological clusters are grown from the seeds by aggregating crystals with at least one side in common with a crystal already in the cluster, and with an energy exceeding another threshold. A topological cluster gives rise to as many PF clusters. Third, the energy of each cell is shared among all PF clusters according to the cell-cluster distance, with an iterative determination of cluster energies and positions. PF clusters are then assembled into PF superclusters, starting from a seed cluster and gathering the presumptive clusters of bremsstrahlung photons and conversions products.
Electron tracks are reconstructed with a dedicated tracking procedure, different from the one used for other charged particles, because electrons lose a larger amount of energy in the tracker and this causes changes in the curvature. The tracker procedure is preceded by a seeding procedure that consists in finding and selecting the two or three first hits in the tracker from which the track can be initiated. Then the track building proceeds iteratively from the track parameters provided in each layer, modelling the electron energy loss with a Bethe-Block function. To maintain good efficiency in the presence of bremsstrahlung, compatibility requirements between the predicted and the found hits in each layer are quite loose. If several hits are compatible with the predicted one, different trajectory candidates are created and developed, with a limit of five candidate trajectories for each layer. At most one missing hit is allowed per each trajectory. Once the hits are collected, the track parameters are estimated with a fit that uses a Gaussian Sum Filter (GSF) [22], instead of the Kalman Filter (KF) [23] used for non-electron tracks.

In the end, a refining procedure is applied to the PF superclusters and some loose requirements are applied to the variables that characterize the geometrical association between the track and the supercluster. Charge estimation is then performed and the final step is the estimation of electron momentum, which relies on a combination of the energy of the supercluster and the momentum estimate of the GSF track.

3.1.2 Muon reconstruction

Muon tracks are initially reconstructed independently in the inner tracker (tracker tracks) and in the outer muon system (standalone-muon tracks). Then the muon reconstruction algorithm combines the information from both subsystems using two approaches:

- Global muons are formed by propagating standalone-muon tracks inward to the inner tracker. If a matching tracker track is found, a global-muon track is fitted combining the hits from the tracker and standalone-muon track, using the Kalman Filter technique;
- Tracker muons are formed by extrapolating tracker tracks outward to the outer muon system, requiring that at least one muon track segment made of hits in the DTs or CSCs matches the extrapolated track. The possibility for tracker muons to have one single matched segment in the muon system makes this algorithm more efficient than global-muon reconstruction at low momentum ($p_\mu \lesssim 5$ GeV).

Then muon candidates found by both algorithms that share the same tracker track are merged into a single candidate. The charge and momentum are taken from the tracker track for muons of $p_\mu^T < 200$ GeV, because the precision of the muon system measurement at low momentum is limited by multiple scattering. For muons with $p_\mu^T > 200$ GeV, charge and momentum are extracted from the combined trajectory fit of the two muon systems.
3.2 Lepton selection

After muon and electrons candidates are reconstructed, they need to pass some selection criteria in order to be used in the $H \rightarrow ZZ^* \rightarrow 4\ell$ analysis. Lepton selection is necessary in order to consider only leptons coming from signal events and reject the ones deriving from backgrounds. Leptons generated by signal events have particular properties which derive from the kinematics of the event, that leptons from the backgrounds have not. Selection cuts are optimized in order to consider in the analysis only leptons by signal events.

The lepton selection involves three main requirements: impact parameter, identification and isolation.

To be considered for the analysis, reconstructed electrons are required to have a transverse momentum $p^e_T > 7$ GeV, while muons are required to have $p^\mu_T > 5$ GeV. A pseudorapidity cut of $|\eta^e| < 2.5$ is applied on electrons, since they are reconstructed within the extent of silicon tracker, while for muons the acceptance cut is $|\eta^\mu| < 2.4$, due to the configuration of the outer muon system.

Reconstructed tracks are then required to point to a common vertex, which is called primary vertex and it is defined as the one that has the largest sum of $p^2_T$ of clusters of associated tracks. Non-prompt muons that originate from in-flight decays of hadrons and cosmic rays, and electrons from photon conversions can be suppressed using the fact that their track does not point to the main primary vertex of the event. A first loose vertex compatibility requirement is applied to electrons and muons as:

$$d_{xy} < 0.5 \text{ cm} \quad \text{and} \quad d_z < 1 \text{ cm} \quad (3.1)$$

where $d_{xy}$ and $d_z$ are the electron or muon impact parameter with respect to the primary vertex in the transverse plane and in the longitudinal direction, respectively. Then a second vertex requirement is applied after defining a three-dimensional impact parameter significance SIP$_{3D}$ as the ratio of the impact parameter of the lepton track (IP$_{3D}$) in three dimensions computed with respect to the primary vertex position, and its uncertainty:

$$\text{SIP}_{3D} = \frac{\text{IP}_{3D}}{\sigma_{\text{IP}_{3D}}} \quad (3.2)$$

The vertex related cut is applied as SIP$_{3D} < 4$, which is a very loose cut. If a track has a larger value of SIP$_{3D}$, it is not associated to the primary vertex.

Leptons that satisfy the above requirements are labelled loose leptons.

The lepton selection then is completed with some additional criteria referred to electron-specific or muon-specific requirements of quality. These criteria aim at selecting actual electrons and muons among all the objects that can be mistakenly reconstructed as electron or muon candidates.

Muon candidates pass identification if they satisfy the requirements of the PF algorithm. PF muon identification proceeds in different steps. First an isolation criterion is applied
to already select muons that have little neighbouring activity. Then non-isolated muon candidates are selected as PF muons if they both include a minimum number of hits in the muon track and satisfy a compatibility criterion of the muon segment and calorimeter deposits.

Electron identification is obtained applying a series of cuts on different observables, which are related to the electron supercluster in the ECAL, to electron track reconstruction and that describe the energy-momentum and geometrical matching between the supercluster and the electron track.

Studies on the lepton selection efficiency were performed for the analysis concerning measurement of Higgs boson properties in the $H \rightarrow ZZ \rightarrow 4\ell$ channel in the context of LHC Run I. Figure 3.2 shows the transverse momentum $p_T$ distributions at generator level for the four leptons from $H \rightarrow ZZ \rightarrow 4\ell$ signal events obtained with MC samples for a mass hypothesis of $m_H = 126$ GeV/$c^2$. The empty histograms are the $p_T$ distributions for generated leptons, while the shaded histograms are the $p_T$ distributions after the lepton selection. These plots show the necessity of selecting also low-$p_T$ leptons in order to improve the sensitivity of the analysis.
Figure 3.2: Distribution of the transverse momentum ($p_T$) for each of the four leptons (ordered in $p_T$) from $H \to ZZ \to 4\ell$ signal events, for a mass hypothesis of $m_H = 126$ GeV/c$^2$. The distributions are obtained using MC signal samples and shown at generator level within $\eta$ acceptance (empty histograms), and for selected events (shaded histograms) in the $4e$ (a), $4\mu$ (b), and $2e2\mu$ (c) channel \[16\].
3.3 Lepton isolation

This is a powerful mean to suppress background from non-prompt or misidentified leptons, coming from the weak decays of hadrons within jets. The lepton isolation is defined as:

\[ I_\ell = \frac{1}{p_T} \left( \sum p_T^{\text{charged}} + \max \left[ 0, \sum p_T^{\text{neutral}} + \sum p_T^\gamma - p_T^{\text{PU}}(\ell) \right] \right) \quad (3.3) \]

and the isolation requirement for leptons is \( I_\ell < 0.35 \). The sums in the definition \(3.3\) are restricted to a volume bounded by a cone of angular radius \( R = 0.3 \) around the lepton direction at the primary vertex. The angular distance between two particles \( i \) and \( j \) is defined as:

\[ \Delta R(i,j) = \sqrt{(\eta^i - \eta^j)^2 + (\phi^i - \phi^j)^2} \quad (3.4) \]

In eq. \(3.3\) the \( \sum p_T^{\text{charged}} \) is the scalar sum of the transverse momenta of charged hadrons originating from the chosen primary vertex of the event. \( \sum p_T^{\text{neutral}} \) and \( \sum p_T^\gamma \) are the scalar sums of the transverse momenta for neutral hadrons and photons, respectively. The isolation variable is particularly sensitive to energy deposits from pileup interactions, therefore a \( p_T^{\text{PU}}(\ell) \) contribution is subtracted (with the constraint for the difference to be non negative), using different techniques for muons and electrons.

- For muons it is defined as \( p_T^{\text{PU}}(\mu) = \Delta \beta \equiv 0.5 \sum_i p_T^{\text{PU},i} \), where \( i \) runs over the momenta of the charged hadron PF candidates not originating from the primary vertex, and the factor 0.5 corrects for the different fraction of charged and neutral particles in the cone.

- For electrons it is defined as \( p_T^{\text{PU}}(e) \equiv \rho A_{\text{eff}} \), where the effective area \( A_{\text{eff}} \) is the geometric area of the isolation cone scaled by a factor that accounts for the residual dependence of the average pileup deposition on the \( \eta \) of the electron, while \( \rho \) is the mean of the \( p_T \) density distribution of neutral particles in the event.

3.4 FSR photon recovery

Leptons from \( Z \) bosons decays can radiate a high-energy photon, a phenomenon called final state radiation (FSR). This fact degrades the accuracy of the information that can be extracted from the four-lepton candidates selected in the analysis. Therefore it is important to identify and collect these photons, associating them to their parent leptons, in order to fully reconstruct the Higgs boson decay system. The FSR recovery algorithms is designed to discriminate FSR photons from the background, due to pileup interactions or initial state radiation. This is done exploiting the particular kinematics of FSR photons, that are irradiated in a direction collinear to the one of their parent lepton, and tend to be isolated from other particles.
FSR photons candidates are considered if they satisfy the requirements of pseudorapidity $|\eta^\gamma| < 2.4$, transverse momentum $p_T^\gamma > 2\text{GeV}$ and isolation $I^\gamma < 1.8$. The relative isolation $I^\gamma$ for photon candidates is computed as:

$$I^\gamma \equiv \frac{1}{p_T^\gamma} \left( \sum p_T^\gamma + \sum p_T^{\text{neutral}} + \sum p_T^{\text{charged}} \right)$$

where the $\sum p_T^\gamma$, $\sum p_T^{\text{neutral}}$ and $\sum p_T^{\text{charged}}$ are the scalar sums of the transverse momenta of photons, neutral and charged hadrons contained inside a cone of radius $R = 0.3$. The contribution from pileup is included in the sums.

Then selected photons are associated to the closest lepton in the event among all those pass both the loose ID and SIP cuts. Photons that do not satisfy the cut $\Delta R(\gamma, \ell)/(E_T^\gamma)^2 < 0.012$ and $\Delta R(\gamma, \ell) < 0.5$ are discarded. If more than one photon is associated to the same lepton, only the one with the lowest $\Delta R(\gamma, \ell)/E_T^\gamma$ is selected.

Since the FSR photons are often located in the isolation cone of their lepton, they tend to make it fail the isolation requirement. Therefore, all selected FSR photons are explicitly subtracted from the isolation sums of all loose leptons in the event, not only of their associated leptons.

### 3.5 Lepton momentum calibration

The determination of the electron momentum relies on a combination from ECAL and tracker, while for muons it involves the tracker and the muon chambers. The calibration of the tracker, the ECAL and the muon system relies on the best knowledge of the detector conditions, but some small discrepancies can remain between data and simulation. Therefore the scale and the resolution of lepton momenta have to be calibrated. This is done in bins of $p_T^\ell$ and $\eta^\ell$, exploiting some well-known di-lepton resonances.

The scale of electrons is calibrated using a $Z \rightarrow e^+e^-$ control sample, correcting the momenta as to align the reconstructed di-electron mass spectrum in the data to that in the MC, and to minimize the width of the distribution, as shown in Figure 3.3. Time-dependent variation of electron momenta may also happen, due to loss of the transparency of ECAL crystals. In order to account this variations, the correction is derived as a function of time. In addition to all this, a smearing of the electron energies is applied in simulation so as to make the $Z \rightarrow e^+e^-$ mass resolution in simulation match that observed in data. In the end, a pseudo-random Gaussian multiplicative factor is applied, with Gaussian parameters varying in bins of $p_T^e$ and $|\eta^e|$.

The scale of muons, instead, is calibrated using a Kalman filter approach [24], using the $J/\psi$ meson and the $Z$ boson decay into muons ($Z \rightarrow \mu^+\mu^-$). This technique corrects the muon track in the silicon tracker for three different effects: small variations of the magnetic field, residual misalignment, and the imperfect modelling of the material. Then,
3.6 Lepton efficiency measurements

The efficiency of reconstructing and selecting signal leptons is an important point of the analysis. It needs to be optimized with respect to background leptons and to be accurately measured both in data and MC in order to correct possible discrepancies. Lepton efficiency measurements are based on the Tag-and-Probe technique described in the next section, and the measurement is done in several bins of $p_T$ and $\eta$.

The efficiency measurement is done in data and MC simulations with the same method, in order to avoid any bias. Therefore it is possible to define a per-lepton efficiency scale factor as:

$$SF_\ell(p_T, \eta) = \frac{\epsilon_{\text{data}}(p_T, \eta)}{\epsilon_{\text{MC}}(p_T, \eta)}$$ (3.6)

Then, considering that for each event four leptons are selected, a per-event data/MC
scale factor is defined as the product of the scale factor of the selected leptons:

\[
\text{SF}_{4\ell} = \prod_{\ell=1}^{4} \text{SF}_{\ell}(p_T^\ell, \eta^\ell)
\]  

(3.7)

which is used to re-weight MC samples event by event.

### 3.6.1 The Tag and Probe technique

The Tag-and-Probe technique, in general, consists in using a known mass resonance (e.g. J/ψ, Υ, Z) to select particles of the desired type, and probe the efficiency of a particular selection criterion on those particles.

For the current analysis a pure sample of \(Z \rightarrow \ell^+\ell^-\) is used, in which, at first, a lepton with very tight selection and isolation requirements is selected, and it is referred to as “tag”. Then, another lepton of same flavour and opposite charge, and as loose selection as possible, is taken for a certain mass window around the mass of the considered resonance, in our case Z mass \(m_Z\). This second lepton is referred to as “probe”. Then, the efficiency of a selection cut is estimated as:

\[
\epsilon = \frac{N_P}{N_P + N_F}
\]

(3.8)

where \(N_P\) is the numbers of probes passing the selection criteria and \(N_F\) is the number of probes failing them.

Despite being linked to tags by the Z resonance, probes are not a perfectly pure source of leptons. This means that a cut-and-count approach may not be reliably enough to estimate the efficiencies. Therefore the di-lepton invariant mass distribution of passing and failing probes are fitted separately with a signal plus background model. This way, taking into account background (Drell-Yan events and fake leptons), the the actual number of passing and failing leptons is provided. Then the efficiency is computed from the ratio of the signal yields of the two lineshapes obtained. The procedure is repeated in different bins of \(p_T^\ell\) and \(\eta^\ell\). The ratio of the efficiencies measured in simulation and data is used to rescale the selection efficiency in the simulated samples, as described above.

### 3.7 Jet reconstruction and selection

The aim of the present thesis is to investigate the secondary Higgs boson production mechanisms, illustrated in Chapter I. Most of these production modes, like VBF and VH, present particle jets in the final state, due to decay of the additional particles produced. The presence of jets, therefore, can be used to divide selected events into categories, needed to improve the analysis sensitivity to the different production mechanisms. Therefore it is important to select and reconstruct particle jets.
Particle jets are reconstructed using the anti-$k_T$ clustering algorithm [26] on PF candidates, with a distance parameter $R = 0.4$, after rejecting the charged hadrons that are associated to a pileup primary vertex. In this analysis, the jets are required to be within $|\eta| < 4.7$ “area” and to have a transverse momentum $p_T > 30$ GeV. Since also electrons and photons can be reconstructed as jets, a jet-cleaning procedure is implemented requiring the selected jets to be geometrically separated from any of the tight leptons passing the SIP and isolation cut computed after FSR correction and from their FSR photons by an angular distance $\Delta R(\text{jet}, \ell/\gamma) > 0.4$.

Since calorimeter response to particles energy is not linear, some jet energy corrections (JEC) are applied to jets in both simulation and data. These corrections are computed for each data taking period and they include a calibration of the jet energy scale (JES) and a smearing of the jet energy resolution in simulation in order to match that observed in data.

For categorization purposes, it is necessary to distinguish whether a jet derives from the hadronization of a bottom quark (b-jet) or not. In fact processes like $t\bar{t}H$ and $ZH$ involve b jets. Therefore, b-jet identification techniques are used. They rely on the fact that B hadrons present in the b jets have a long lifetime and therefore decay a few millimetres away from the primary vertex. Hence, in the reconstruction the tracks result displaced and the secondary vertex to which they point can be reconstructed. The b-tagging algorithm used in this analysis is the Combined Secondary Vertex algorithm (CSVv2) [26], which combines information about the impact parameter significance of the tracks, the secondary vertex and the jet kinematics.

The imperfect modelling of the distributions of the b tagging variables in data causes some discrepancies between data and MC in the b-jet identification efficiency and misidentification probability. In order to correct these differences, scale factors are applied to the MC and they are defined as the ratio of the efficiencies of tagging actual b jets in data and simulation:

$$ SF = \frac{\epsilon_{\text{data}}}{\epsilon_{\text{MC}}} $$  \hspace{1cm} (3.9)

These corrections are provided by CMS as a function of the jet $p_T$ and $\eta$.

### 3.8 Event selection

The $H \to ZZ \to 4\ell$ channel is characterized by a very clean signature of four isolated leptons originating from the same vertex and compatible with the decay of a pair of Z bosons. The process has a low cross section but with an appropriate event selection it is possible to achieve an excellent signal-to-background ratio.

The selection of events relies on two main part: the on-line trigger selection requirements and the signal region selection.
3.8.1 Trigger requirements

The trigger is the system that takes care of the on-line selection of events to be recorded for further off-line reconstruction and analysis. The LHC bunch crossings occur at a rate of 40 MHz but only a small fraction of the bunch crossings contain event of interest to the considered analysis. Therefore it is necessary to accurately select the events of interest, and this is done with the trigger. The trigger selects in real time the small fraction of collision events that are relevant to physics analysis and store them for later analysis.

The CMS trigger system relies on two successive levels:

- The first level, the Level-1 trigger (L1), performs a fast readout of the detector with a limited granularity, selecting events that contain distinctive detector signals such as ionization deposits consistent with a muon, or energy clusters compatible with an electron or a photon, with suitable momentum or energy thresholds.

- The second level, the High Level Trigger (HLT), is implemented in software and performs a full readout of the CMS detector. Events are reconstructed with similar algorithms as those used in the off-line analysis, with progressive selection steps in order to allow more sophisticated and time consuming algorithms to be applied over a smaller fraction of events. All main classes of physics objects can be reconstructed at HLT, like electrons, muons or photons, and on these objects are applied specific selection criteria in order to keep the rate under control and select the subset of events relevant to subsequent data analysis.

For the $H \to ZZ^* \to 4\ell$ analysis, a dedicated collection of HLT paths has been designed and optimized to cover the phase space of the $4\ell$ signal. To reproduce the selection efficiency in the data, a simulated trigger selection is applied to Monte Carlo samples as well.

A HLT path consists of several steps made by software modules. Each module performs a well defined task, such as unpacking raw data coming form detectors or reconstructing physics objects (muons, electrons or jets). All these intermediate decisions filter the collected data selecting only events with required characteristics. There are several HLT path and each of them selects a particular kind of events. In the current analysis the HLT paths used select events which present in the final state a certain number of electrons or muons since this analysis is focused on the $H \to ZZ \to 4\ell$ decay channel. Table 3.1 lists the HLT paths used for 2016 analysis. They are grouped according to the kind of event selected: the DiEle triggers select events which present two electrons in the final state, the TriEle select those events with three electrons, DiMuon those with two muons, TriMuon the ones with three muons, MuEle those events with at least one muon and one electron, SingleMuon those with one single muon in the final state, and SingleElectron the events with one electron in the final state. Each HLT in these groups then requires the event to satisfy particular conditions regarding the transverse impulse of particles or the particle identification from the sub-detector systems. For example the HLT HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ selects events with two electrons...
one with $p_T > 17 \text{ GeV}/c$ and the other with $p_T > 12 \text{ GeV}/c$ that passed the identification in the ECAL, in the tracker detector and the isolation requirements.

For the current analysis any event where at least one of the dedicated HLT paths is fired is selected, without matching the three possible final state ($4e$, $4\mu$, $2e2\mu$) to a particular path. This is done for taking into account the possible decays due to different production modes. For example the triple-muon path is not connected only to the $4\mu$ final state because an event produced with VH and with final state $2e2\mu$ can present an additional muon coming from the decay of the associated Z or W boson.

The $H \rightarrow ZZ \rightarrow 4\ell$ analysis relies on five different primary datasets (PDs), DoubleEG, DoubleMuon, MuEG, SingleElectron and SingleMuon, each of which is result of a certain collection of HLT paths. In order to avoid duplicate events from different primary datasets, events are taken:

- from DoubleEG if they pass the DiEle or TriEle triggers;
- from DoubleMuon if they pass the DiMuon or TriMuon triggers and fail the DiEle and TriEle triggers;
- from MuEG if they pass the MuEle triggers and fail the DiEle, TriEle, DiMuon and TriMuon triggers;
- from SingleElectron if they pass the SingleElectron trigger and fail all the above triggers;
- from SingleMuon if they pass the SingleMuon trigger and fail all the above triggers.

The primary datasets used for 2016 analysis are listed in Table 3.1 together with their associated HLT paths.

Using the presented trigger requirements, the selection of candidate events is performed as described in the following section.

### 3.8.2 Candidate selection

The candidate event selection relies on the choice of one four-lepton candidate ($ZZ$ candidate) per event. The four-lepton candidates are built from the so-called selected leptons, which are the reconstructed leptons that pass the impact parameter, identification and FSR-corrected isolation requirements described in previous sections. From this starting point, the selection of the four-lepton candidates proceeds according to the following steps.

**Z candidates** are defined as pairs of selected leptons of opposite charge and matching flavour ($e^+e^-$, $\mu^+\mu^-$) that satisfy $12 < m_{\ell\ell(\gamma)} < 120 \text{ GeV}/c^2$, where the Z candidate mass includes the selected FSR photons if any.
**ZZ candidates** are defined as pairs of non-overlapping Z candidates. The Z candidate with reconstructed mass $m_{\ell\ell}$ closest to the nominal Z boson mass is called $Z_1$, and the other one $Z_2$. The selected FSR photons are included in invariant mass computations. ZZ candidates are required to satisfy the following list of requirements:

- **Ghost removal**: there must be an angular distance $\Delta R(\eta, \phi) > 0.02$ between each of the four leptons. This requirement protects from “ghost tracks” made by a fraction of hits in the tracker detector;

- **Lepton $p_T$**: two of the four selected leptons should pass $p_{T,1} > 20 \text{ GeV}/c$ and $p_{T,2} > 10 \text{ GeV}/c$, in order to assure a tighter selection than the simple trigger selection;

- **QCD suppression**: all four opposite-sign pairs that can be built with the four leptons (regardless of lepton flavour) must satisfy $m_{\ell\ell} > 4 \text{ GeV}/c^2$. This cut suppresses pairs of leptons from cascade decays, which may be different-flavour and are found to broadly peak at very low invariant masses. In this case, selected FSR photons are not used in computing $m_{\ell\ell}$, since a QCD-induced low mass di-lepton (eg. $J/\psi$) may have photons nearby (eg. from a $\pi_0$).

- **$Z_1$ invariant mass**: $m_{Z_1} > 40 \text{ GeV}/c^2$;

- **Wrong pairing suppression**: defining $Z_a$ and $Z_b$ as the mass-sorted alternative pairing Z candidates ($Z_a$ being the one closest to the nominal Z boson mass), the ZZ candidate is excluded if $m_{Z_b} < 12 \text{ GeV}/c^2$ while $m_{Z_a}$ is closer to the nominal Z boson mass than $m_{Z_1}$ is (i.e. NOT($|m_{Z_a} - m_Z| < |m_{Z_1} - m_Z|$ AND $m_{Z_b} < 12 \text{ GeV}/c^2$)). This cut discards $4\mu$ and $4e$ candidates where the alternative pairing are an on-shell Z + low-mass $\ell^+\ell^-$ resonance.

- **Four-lepton invariant mass**: $m_{4\ell} > 70 \text{ GeV}/c^2$.

Events containing at least one selected ZZ candidate form the signal region.

### 3.8.3 Choice of the best ZZ candidate

Some of the events selected in the signal region may have more than one ZZ candidate selected. This can happen when more than four leptons are selected, like for example in VH or $t\bar{t}H$ events where the associated W, Z or $t\bar{t}$ particles decay leptonically. One candidate then has to be chosen.

In the 2016 analysis the choice of the best ZZ candidate is performed after all kinematic cuts have been applied. In this analysis if more than one ZZ candidate survives the above selection, the “most signal-like” is chosen, i.e. the one with the highest value of the kinematic discriminant $D_{bkg}^\text{kin}$ (which will be defined in Chapter 5). In the case when four same-flavour leptons are present in the event, two candidates can be made, which differ only by the values of the $Z_1$ and $Z_2$ masses. Since they have identical values of $D_{bkg}^\text{kin}$, in this case the ZZ candidate with the $Z_1$ mass closest to the nominal Z boson mass is chosen.
Figure 3.4: Selection efficiency within the geometrical acceptance for the SM Higgs boson signal as a function of $m_H$ in the three final states for gluon fusion production. Points represent efficiency estimated from full CMS simulation; lines represent a smooth polynomial curve interpolating the the SM Higgs boson signal as a function of $m_H$ points, used in the Run I analysis [16].

### 3.8.4 Event selection efficiency

The efficiency of event selections after the full event selection is shown in Figure 3.4 for the three final states as a function of Higgs boson mass hypothesis. This study of the selection efficiency as a function of Higgs boson mass was performed in the Run I $H \rightarrow ZZ \rightarrow 4\ell$ analysis [16]. If the Higgs boson mass is $m_H > 200 \text{ GeV}/c^2$, it produces two on-shell Z bosons which decay to leptons with $p_T \sim 45 \text{ GeV}/c$. For these leptons the selection efficiency is quite good. In particular for high-$p_T$ muons the efficiency is $\epsilon_\mu > 98\%$ which implies a total efficiency of about $\sim 87\%$ on the plateau in the plot. For electrons with high $p_T$ the efficiency is $\epsilon_e > 90\%$ and this implies a total efficiency of about 70%. For a lower Higgs boson mass hypothesis, it is possible to see from the plot that the selection efficiency drops. This is due to the Higgs boson decaying in one Z boson on-shell and one Z boson off-shell, which leads to leptons with lower $p_T$, whose selection efficiency is lower.

Once the events are reconstructed and the signal region is defined selecting the ZZ candidates, in order to extract the Higgs boson properties, it is necessary to perform a statistical analysis. For this purpose a statistical model has to be built in order to extract and interpret results from selected data. In the next chapter the statistic model used and the inputs needed in order to extract the results are presented.
<table>
<thead>
<tr>
<th>HTL path</th>
<th>PDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiEle</td>
<td>DoubleEG</td>
</tr>
<tr>
<td>HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ</td>
<td></td>
</tr>
<tr>
<td>HLT_DoubleEle33_CaloIdL_GsfTrkIdVL</td>
<td></td>
</tr>
<tr>
<td>TriEle</td>
<td>DoubleEG</td>
</tr>
<tr>
<td>HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL</td>
<td></td>
</tr>
<tr>
<td>DiMuon</td>
<td>DoubleMuon</td>
</tr>
<tr>
<td>HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL</td>
<td></td>
</tr>
<tr>
<td>TriMuon</td>
<td>DoubleMuon</td>
</tr>
<tr>
<td>HLT_TripleMu_12_10_5</td>
<td></td>
</tr>
<tr>
<td>MuEle</td>
<td>MuEG</td>
</tr>
<tr>
<td>HLT_Mu8_TrkIsoVVL_Ele17_CaloIdL_TrackIdL_IsoVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu17_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Mu8_DiEle12_CaloIdL_TrackIdL</td>
<td></td>
</tr>
<tr>
<td>HLT_DiMu9_Ele9_CaloIdL_TrackIdL</td>
<td></td>
</tr>
<tr>
<td>SingleElectron</td>
<td>SingleElectron</td>
</tr>
<tr>
<td>HLT_Ele25_eta2p1_WPTight_Gsf</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele27_WPTight_Gsf</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele27_eta2p1_WPLoose_Gsf</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele32_eta2p1_WPTight_Gsf</td>
<td></td>
</tr>
<tr>
<td>SingleMuon</td>
<td>SingleMuon</td>
</tr>
<tr>
<td>HLT_IsoMu20</td>
<td></td>
</tr>
<tr>
<td>HLT_IsoTkMu20</td>
<td></td>
</tr>
<tr>
<td>HLT_IsoMu22</td>
<td></td>
</tr>
<tr>
<td>HLT_IsoTkMu22</td>
<td></td>
</tr>
<tr>
<td>HLT_IsoMu24</td>
<td></td>
</tr>
<tr>
<td>HLT_IsoTkMu24</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Trigger paths used in 2016 collision data and their corresponding primary datasets.
Chapter 4

The statistical apparatus

Once the Higgs boson signal has been observed in 2012, further studies on this particle have been performed in order to investigate its nature. With the increasing of the integrated luminosity in Run II, the larger amount of data available makes other properties of this particle accessible, like its sub-dominant production modes. The goal of the current analysis is to study this secondary production mechanisms and to eventually measure their cross section.

This chapter describes the statistical model used to extract and interpret results from the events selected as described in the previous chapter.

4.1 A statistical model

Before performing a statistical analysis, it is necessary to define a model for the data. A statistical model is a set of probability distributions which describe data given some certain parameters:

\[ P(\text{data} \mid \text{parameters}) \]  \hspace{1cm} (4.1)

The motivation of building the statistical model is the possibility to generate pseudo-datasets according to the probability distributions which describe data. This feature is necessary to perform the statistical analysis that allows the significance of an excess in data to be quantified, or the presence of a signal to be excluded. This method was used also in 2012 for the Higgs boson discovery.

Once the excess in data is observed, the same statistical model can be used to estimate the parameters of the model. This is done performing the analysis fixing the data to the values observed in the experiment and leaving the parameters floating, in order to determine their values. This is the case of measuring Higgs boson properties, once the particle has been observed.
4.1.1 The likelihood function

A key point for building the model is the definition of the likelihood function, which contains all the information of the statistical model. The likelihood function is defined as the Probability Density Function (PDF) that characterizes the set of experimental observables, given the parameters of the model. If one experiment measures the values $x_1, ..., x_n$ of $n$ random variables and the model depends on $m$ parameters $\Theta_1, ..., \Theta_m$, the likelihood function $L$ is defined as:

$$L_0(x_1, ..., x_n|\Theta_1, ..., \Theta_m) = f(x_1, ..., x_n|\Theta_1, ..., \Theta_m)$$ (4.2)

where $f$ is the PDF of the random variables $x_1, ..., x_n$.

If there are $N$ events in which the $n$ random variables $x_1, ..., x_n$ are measured, the PDF of the total sample $\vec{x} = \{(x^1_1, ..., x^n_1), ..., (x^N_1, ..., x^N_n)\}$ can be considered. If the $N$ events are independent of each other, the likelihood function of the sample of $N$ events can be written as the product of the PDFs corresponding to the measurement of each single event:

$$L_0(\vec{x}|\vec{\Theta}) = \prod_{i=1}^{N} f(x^i_1, ..., x^i_n|\Theta_1, ..., \Theta_m)$$ (4.3)

where $\vec{\Theta} = (\Theta_1, ..., \Theta_m)$.

If also the number of events $N$ is a random variable and follows the Poisson distribution

$$p(N | \vec{\Theta}) = \frac{e^{-\nu(\vec{\Theta})}\nu(\vec{\Theta})^N}{N!}$$ (4.4)

whose average $\nu$ may depend on the $m$ parameters of the model, the extended likelihood function is defined as:

$$L_{\text{ext}}(\vec{x}|\vec{\Theta}) = p(N | \vec{\Theta}) \cdot L_0(\vec{x}|\vec{\Theta}) = \frac{e^{-\nu(\vec{\Theta})}\nu(\vec{\Theta})^N}{N!} \prod_{i=1}^{N} f(x^i|\vec{\Theta})$$ (4.5)

The likelihood function contains the probability functions which characterise the experimental observables which are considered in the analysis, and therefore contains all the informations necessary to perform the analysis, generating the pseudo-experiments for hypothesis testing (Section 4.2) or performing fits to measure the parameters of the model (Section 4.4).

---

1Considering a random variable $x$, the Probability Density Function $f(x)$ is defined by

$$P(x \in [x, x + dx]) = f(x)dx$$

where $P$ is the probability of finding the value $x$ in the interval $[x, x + dx]$. 
4.1.2 Incorporating systematic uncertainties

In order to correctly define the statistical model, it is necessary to consider some parameters that are not of direct interest to the considered issue. For example, if the issue is to determine the yield of a signal peak, other parameters have to be derived from data, like the experimental resolution which determines the peak width or the detector efficiencies that are necessary to determine the amounts of signal and backgrounds. These parameters are referred to as nuisance parameters. In some cases, they cannot be derived from the same considered data sample, and they have to be estimated with external measurements. The uncertainty on their determination will reflect on the parameters of interest of the model.

Uncertainties due to the propagation of the imperfect knowledge of nuisance parameters give raise to systematic uncertainties, as they reflect a possible deviation of a certain variable from its real value. Therefore in the model each independent source of systematic uncertainty is assigned a nuisance parameter $\theta_i$. Their full set is denoted as $\theta$, and the $m$ parameters of the model, $\Theta = \theta_1, ..., \theta_m$, are functions of these nuisance parameters: $\Theta(\theta)$.

Since the nuisance parameters are constrained by other measurements, to take into account the probability to measure a value $\tilde{\theta}_i$ for the $i$-th parameter, given its true value $\theta_i$, the PDF $p_i(\tilde{\theta}_i|\theta_i)$ associated to each nuisance parameter is considered. The PDF for all the nuisance parameters is $p(\tilde{\theta} | \theta)$.

Hence, the likelihood function of the model, incorporating the nuisance parameters, becomes:

$$L(\bar{x}|\Theta(\theta)) = p(\tilde{\theta} | \theta) \cdot L_{\text{ext}}(\bar{x}|\Theta(\theta)) = p(\tilde{\theta} | \theta) \cdot \frac{e^{-\nu(\tilde{\Theta}(\theta))} \nu(\tilde{\Theta}(\theta))^N}{N!} \prod_{i=1}^{N} f(x_i|\Theta(\theta)) \quad (4.6)$$

4.2 Hypothesis testing

Once the model has been defined, the statistical analysis can be performed on the collected data. In order to illustrate the function of the statistical model and of the likelihood function defined in the previous section, in this section is considered the simplest and most common case of statistical analysis which is hypothesis testing. In fact one of the key tasks in most of physics measurements is to discriminate between two hypotheses on the basis of the observed experimental data. A typical example in physics is the identification of the type of a considered particle, for example muon instead of pion, and this is done on the basis of the measurement of discriminating variables provided by a particle-identification detector.

In statistical literature when two hypotheses are present, the first one is called null hypothesis $H_0$, while the second one is labelled alternative hypothesis $H_1$.

Let’s assume that the result of a measurement of $n$ variables is $\bar{x} = (x_1, ..., x_n)$, which is a random variable distributed according to a PDF $f(\bar{x})$ that is in general different
under the hypotheses $H_0$ and $H_1$. A measure of whether the observed data sample better agrees with the hypothesis $H_0$ or rather with $H_1$, can be given by a test statistic. A test statistic, $t(\bar{x})$, is a function of the measured sample $\bar{x}$, and it in turn is a random variable distributed according to a PDF $g(t)$, that can be derived from the PDF of the observed sample $\bar{x}$. The test statistic must be defined in order to have different distributions ($g(t|H_0)$ and $g(t|H_1)$) under the two different hypotheses $H_0$ and $H_1$ (see Figure 4.1).

The test statistic is defined to discriminate the two hypotheses. Having the statistical model, the expected distribution of $g(t)$ under the two hypotheses ($g(t|H_0)$ and $g(t|H_1)$) can be reproduced by generating pseudo-experiments from the PDFs contained in the likelihood function of the model. The knowledge of these distributions allows to derive the probability of a given value of $t$ to be observed under either hypothesis. In hypothesis testing, a value, $t_{cut}$, is set corresponding to the desired confidence level of the test, based on the distribution of the test statistic for the null hypothesis.

On the basis of the result of the measurement $x_{obs}$, the test $t$ is evaluated in the observed value, defining the measured value of the test statistic:

$$\hat{t} = t(x_{obs})$$  \hspace{1cm} (4.7)

This value is then compared with the expected distributions, and it is possible to deter-
mine if the null hypothesis is a good description for the collected data. Therefore:

\[
\begin{align*}
\text{if } \hat{t} \leq t_{\text{cut}} & \quad \text{H}_0 \text{ is accepted} \\
\text{if } \hat{t} > t_{\text{cut}} & \quad \text{H}_0 \text{ is rejected}
\end{align*}
\]  

(4.8)

at the predefined confidence level.

The test statistic is usually built as a ratio of two likelihood functions, one evaluated for the observed data sample under the hypothesis \( \text{H}_0 \) and the other under \( \text{H}_1 \):

\[
\lambda(\vec{x}) = \frac{L(\vec{x}|\hat{\Theta}(\theta), \text{H}_0)}{L(\vec{x}|\hat{\Theta}(\theta), \text{H}_1)}
\]  

(4.9)

This choice is suggested by the Neyman-Pearson lemma \[28\] which states that a test defined as a likelihood ratio allows the best discrimination between two hypotheses to be achieved.

Considering the the Wilks’ theorem \[29\] which states that for a model with \( m \) parameters of interest the distribution of \(-2\Delta \ln L\) approaches a \( \chi^2 \) distribution with \( m \) degrees of freedom in the limit of a large data sample, the test statistic can be defined as:

\[
t(\vec{x}) = -2 \ln \lambda(\vec{x}) = -2 \ln \left( \frac{L(\vec{x}|\hat{\Theta}(\theta), \text{H}_0)}{L(\vec{x}|\hat{\Theta}(\theta), \text{H}_1)} \right) = -2\Delta \ln L
\]  

(4.10)

The presence of nuisance parameters broads the distribution of the test statistic and this fact brings to a loss of information.

An example of hypothesis testing is the measurement of the Higgs boson spin-parity \( J^p \) performed in the \( \text{H} \rightarrow \text{ZZ} \rightarrow 4\ell \) in the LHC Run I \[16\]. As shown in Figure 4.2 the distributions for the two spin-parity hypotheses (\( J^p = 0^{+} \) and \( J^p = 0^{-} \)) were obtained by generating pseudo-datasets. The test statistic, then, was evaluated from the observed data, and the obtained value was compared with these distributions, showing the agreement of observed data with the \( J^p = 0^{+} \) hypothesis.

### 4.3 Quantifying an excess

With the statistical model defined and the hypothesis testing method illustrated, it is possible to quantify the significance of an excess over the background-only expectation. This can be done defining a test statistic which discriminates between:

- the null hypothesis \( \text{H}_0 \), which states that the observed data are due only to background events (\( b \));

- the alternative hypothesis \( \text{H}_1 \), which states that the observed data are due to background plus signal events (\( b + s \)).

In the \( \text{H} \rightarrow \text{ZZ} \rightarrow 4\ell \) analysis a particular parametrization of signal yield is used to quantify an excess. This method was developed in the LHC Run I by the ATLAS

48
Figure 4.2: Distribution of a test-statistic $t = -2 \ln (\mathcal{L}_0^- / \mathcal{L}_0^+)$ of the pseudoscalar boson hypothesis ($J^p = 0^-$) tested against the SM Higgs boson hypothesis ($J^p = 0^+$). Distributions for the SM Higgs boson are represented by the yellow histogram and for the alternative $J^p$ hypotheses by the blue histogram. The red arrow indicates the observed value [16].

and CMS Collaborations in the context of the LHC Higgs Combination Group [30]. It introduces a new parameter: the signal strength modifier $\mu$ [31]. It is defined as the measured cross section normalized to the expectation for the SM Higgs boson:

$$\mu = \frac{\sigma}{\sigma_{SM}}$$  \hspace{1cm} (4.11)

This parameter is used to define the expected number of events $\nu$, introduced in the likelihood function eq[4.6], as:

$$\nu(\vec{\Theta}(\theta)) = \mu \cdot s + b$$  \hspace{1cm} (4.12)

where the expected yields for background $b$ and for Higgs boson signal $s$ are both functions of the other parameters of the model and of the nuisance parameters $\theta$ contained in the model:

$$b = b(\vec{\Theta}(\theta))$$

$$s = s(\vec{\Theta}(\theta))$$  \hspace{1cm} (4.13)
The hypothesis $H_0$ corresponding to the presence of background only requires $\mu = 0$, while the hypothesis $H_1$ which corresponds to the presence of signal plus background allows any non-null positive value of $\mu$.

With this parametrization, the PDF $f(x^i|\Theta(\theta))$ of the likelihood function eq.4.6 can be written as the superposition of two components, one for the signal and another for the background, weighted by the expected signal and background fractions:

$$f(x^i|\Theta(\theta)) = \frac{\mu_s}{\mu_s + b} f_s(x^i|\Theta(\theta)) + \frac{b}{\mu_s + b} f_b(x^i|\Theta(\theta))$$  \hspace{1cm} (4.14)

In this case the likelihood function eq.4.6 becomes:

$$L_{s+b}(\bar{x}|\mu, b, \theta, \Theta(\theta)) = p(\bar{\theta}|\theta) \cdot \frac{e^{-\mu_s b}}{N!} \prod_{i=1}^{N} \left( \mu_s f_s(x^i|\Theta(\theta)) + b f_b(x^i|\Theta(\theta)) \right)$$  \hspace{1cm} (4.15)

where $p(\bar{\theta}|\theta)$ is the PDF for all the nuisance parameter.

Under the background only hypothesis, the signal strength modifier is null $\mu = 0$, and the likelihood in this case is:

$$L_b(\bar{x}|b, \Theta(\theta)) = p(\bar{\theta}|\theta) \cdot \frac{e^{-b}}{N!} \prod_{i=1}^{N} b f_b(x^i|\Theta(\theta))$$  \hspace{1cm} (4.16)

Therefore it is possible to build a test statistic in order to discriminate between the two hypotheses $H_0$ and $H_1$ as a likelihood ratio:

$$q_0 = -2\ln \frac{L_b(x|b, \Theta(\theta))}{L_{s+b}(x|\mu, b, \Theta(\theta))}$$  \hspace{1cm} (4.17)

where the numerator and the denominator are computed with the values of $\mu$ and of nuisances parameters $\theta$ which maximize them. With this definition $q_0$ is positive for a signal-like excess ($\mu > 0$), while in absence of an excess ($\mu = 0$) $q_0$ becomes 0.

The significance of an excess can be quantified in terms of the **p-value**, defined as the probability to obtain a value of the statistic test $q_0$ as large as the one observed in experimental data under the background-only hypothesis:

$$p_0 = \mathcal{P}(q_0 \geq q_0^{\text{obs}} | b)$$  \hspace{1cm} (4.18)

The p-value can be extracted by comparing the observed value of $q_0^{\text{obs}}$ with the expected distribution of $q_0$, which can be obtained by generating pseudo-experiments according to the statistical model defined.
Chapter 4. The statistical apparatus

The p-value is usually converted in the significance $Z$ of the excess through the Gaussian one-sided tail integral:

$$p_0 = \int_{Z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$$

(4.19)

as shown in Fig.4.3. The conventional $Z = 5\sigma$ threshold for claiming a discovery corresponds to a p-value of $2.8 \times 10^{-7}$.

The local p-value $p_0$ characterizes the probability of a background fluctuation resembling a signal-like excess, for a given value of the other parameters of the model $\bar{\Theta}(\theta)$, among which there is the Higgs boson mass.

The probability for a background fluctuation to be at least as large as the observed maximum excess anywhere in a given mass range is called **global p-value**, which can be evaluated by generating pseudo-datasets for different Higgs boson mass hypotheses.

### 4.4 Measuring parameters

Determining the significance of an excess is part of the process that leads to a discovery. Since the Higgs boson was discovered in 2012, the goal of the statistical analysis is measuring its properties.

In particular, the signal strength modifier $\mu$, defined in eq.4.11 is, together with the boson mass, the first property to be measured. If the measure is focused on the signal strength, it represents the single parameter of interest of the model, therefore the other parameters of the model $\bar{\Theta}(\theta)$, like the Higgs boson mass or width, are fixed to their values computed in the theory or measured in other experiments.
Using the likelihood function eq.[4.6] of the defined model, with \( \mu \) as parameter of interest of the model, it is possible to define a test statistic as a likelihood ratio:

\[
q = -2\Delta \ln \mathcal{L} = -2 \ln \frac{\mathcal{L}(\vec{x} | \hat{\mu}, \hat{\theta}_\mu)}{\mathcal{L}(\vec{x} | \hat{\mu}, \theta)}
\]

(4.20)

where \( \hat{\theta} \) and \( \hat{\theta}_\mu \) are the values of nuisances parameters that maximize the likelihood functions at the denominator and at the numerator, respectively, while \( \hat{\mu} \) is the best fit value for the parameters of interest \( \mu \). At the numerator the fit is done fixing \( \mu \) and fitting the nuisance parameters \( \theta \), while at the denominator both \( \mu \) and \( \theta \) are fitted.

The results of the measurement are given as central values of the parameter \( \mu \) with 68% CL intervals, and displayed graphically as scans of \(-2\Delta \ln \mathcal{L}\).

Expected results can also be provided for some nominal values of the parameters. This would in principle require to generate a large number of pseudo-experiments and determine their median outcome, but a good approximation is provided by the Asimov dataset[32], which is a single representative dataset where the observed rates and distributions coincide with predictions under the nominal values of nuisance parameters.

### 4.5 Inputs for the statistical model

The statistical model is composed of a certain number of inputs so that it can be used to generate the pseudo-datasets needed for the statistical analysis. These inputs describe the expected behaviour for the signal and the backgrounds considered in the analysis, and they enter in the defined model as parameters of the PDFs used to generate pseudo-experiments. The inputs needed for the analysis are:

- The number of **expected events** (expected yield), which are the normalization for the signal and for backgrounds;
- The invariant **mass lineshape**, that describes how signal and backgrounds events are distributed in the mass range considered;
- The sources of **systematic uncertainty** and their magnitude, which enter in the model as nuisance parameters.

The expected invariant mass lineshapes are necessary in order to perform a so called **shape analysis**. In this kind of analysis, instead of searching for or study an excess in data relative to the expected signal and background yields in a given mass window (counting experiment), also the expected mass lineshape of a signal-like peak is considered.

In a counting experiment the observed event rate in a certain mass window is compared with the expected yield for a background hypothesis. If an excess of events is observed in data with respect to expected yields, its significance is computed only considering the
numerical excess of signal over background, without looking at the distribution of events inside the considered mass window.

In a shape analysis the observed data are compared with the expected yields like in the counting experiment, but also their invariant mass distribution is considered and compared with the expectation: a scan is performed in a given mass range, testing the expected signal shape with observed data for different values of the resonance mass. Furthermore the shape analysis allows the computation of the excess significance and the measurement of peak properties to be performed with more sensitivity with respect to a counting experiment.

The expected lineshapes are evaluated from the MC simulations, as the expected yields, and since they are analytical functions defined by certain parameters, in statistical model there contains some systematic uncertainties which account the possible variations of these parameters.

The following chapter will present the analysis performed with 2016 data using the statistical model described in this chapter, first describing how the inputs of the model are computed for the $H \rightarrow ZZ \rightarrow 4\ell$ 2016 analysis and then presenting the results obtained.
Chapter 5

Measurement of Higgs boson properties with 2016 data

Once the events have been selected as described in Chapter 3, they can be used for the analysis, whose goal is to measure Higgs boson properties. In particular, thanks to the amount of data available, this analysis is built to investigate the Higgs boson sub-dominant production modes (Section 1.4.1): Vector Boson Fusion (VBF), associate production with a vector boson (VH) or with a top-antitop quark pair (ttH). Investigating the secondary production mechanisms allows to study the Higgs boson couplings with different particles. To quantify the Higgs boson couplings it is necessary to measure their respective cross sections.

The production cross section measurement can be performed both inclusively, i.e. considering together all the modes in which the Higgs boson can be produced, and per production mechanism. The measurement of the cross section dividing per production can give information on the Higgs boson couplings.

In order to perform the measurement of the Higgs boson cross section, a signal strength modifier $\mu$ (Eq. 4.1) is introduced as scale of the expected Standard Model cross section:

$$\sigma = \mu \cdot \sigma_{SM} \quad (5.1)$$

This variable can be measured both using the whole available amount of data (and this gives information on the inclusive Higgs boson cross section) and dividing data in categories enriched in the different production mechanisms, in order to access the corresponding cross-sections. The event categorization used in the analysis will be presented in Section 5.3.

In order to extract the results from collected data, a statistical analysis has to be performed and the statistical model used in the 2016 $H \to ZZ \to 4\ell$ analysis is the one described in the previous Chapter 4. The current Chapter starts with the description of the data and Monte-Carlo samples used in the analysis. The estimation of the input of the statistical model is presented in Section 5.4. Finally the results obtained from the statistical analysis which have been submitted for the publication in [1].


<table>
<thead>
<tr>
<th>Process</th>
<th>$\sigma \times BR$ [fb]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$gg \to H \to ZZ \to 4\ell$</td>
<td>12.18</td>
</tr>
<tr>
<td>$qq \to Hqq \to ZZqq \to 4\ell qq$</td>
<td>1.044</td>
</tr>
<tr>
<td>$qq \to W^+H \to W^+ZZ \to 4\ell + X$</td>
<td>0.232</td>
</tr>
<tr>
<td>$qq \to W^-H \to W^-ZZ \to 4\ell + X$</td>
<td>0.147</td>
</tr>
<tr>
<td>$q\bar{q} \to ZH \to ZZZ \to 4\ell + X$</td>
<td>0.668</td>
</tr>
<tr>
<td>$gg \to t\bar{t}H \to t\bar{t}ZZ \to 4\ell$</td>
<td>0.393</td>
</tr>
</tbody>
</table>

Table 5.1: List of signal Monte Carlo samples and their corresponding production cross sections at 13 TeV, times relevant branching fractions.

5.1 Data and simulated samples

The data sample considered in this analysis is the proton-proton collision data sample recorded by the CMS detector in 2016, corresponding to an integrated luminosity of 35.9 fb$^{-1}$. Collision events are selected by high-level trigger algorithms that require the presence of leptons passing loose identification and isolation requirements, described in Section 3.8.

The Monte Carlo samples are used in this analysis for several purposes, as to estimate backgrounds, optimize the event selection, and evaluate the acceptance and systematic uncertainties.

The description of the SM Higgs boson signal is obtained at Next-to-Leading-Order (NLO) in perturbative Quantum Chromodynamics (pQCD) with the powheg 2.0 generator [33] for the five main production modes: ggH, VBF, WH, ZH and t\bar{t}H. The cross sections for the various signal processes are taken from [8], while the default set of parton distribution functions used for the generation at parton level is taken from [34].

The decay of the Higgs boson to four leptons is modelled with jhugen 7.0.2 [35] which properly accounts for interference effects associated with permutations of identical leptons in the $4e$, $4\mu$ and $4\tau$ final states. Adding $2e2\mu$, $2e2\tau$, and $2\mu2\tau$, six final states in total are included at generator level, even if in the analysis reconstructed tau leptons are not considered. However, because of the existence of leptonic decays of $\tau$ leptons, a small amount of events involving $\tau$ pairs are reconstructed as $4\ell$ ($\ell = e, \mu$) events. In the case of ZH and t\bar{t}H production modes, the Higgs boson is also allowed to decay as $H \to ZZ \to 2\ell 2X$, where $X$ stands for either a quark or a neutrino. In this way also events where two leptons originate from the decay of the associated Z boson or top quarks are taken into account. The list of signal samples and their cross sections are shown in Table 5.1.

Signal samples are generated for five mass points (120, 124, 125, 126, and 130 GeV/$c^2$) in order to parametrise the analysis expectations, to be used as inputs of the statistical model, as a function of the Higgs boson mass $m_H$.

The simulated background samples reproduce the two sources of irreducible background
considered in the current analysis: the production of ZZ pairs through quark-anti-quark annihilation ($q\bar{q} \rightarrow ZZ$), and the production of ZZ pairs through gluon fusion ($gg \rightarrow ZZ$). The reducible background $Z+X$, instead, is estimated with data-driven techniques (Section 5.4.1).

The $q\bar{q} \rightarrow ZZ$ background MC is generated at the Next-to-leading-Order (NLO) in perturbative Quantum Chromodynamics (pQCD) with the powheg 2.0 generator, with the same settings as signal samples. The differential cross section necessary for the MC generation, instead, has been computed at Next-to-Next-to-leading-Order (NNLO) [36], but it is not yet available for the event generator at partonic level. Therefore the MC events have to be corrected for higher level order corrections and this is done weighting each MC event with a constant factor, the so called $k$-factor. To the generated $q\bar{q} \rightarrow ZZ$ sample events, NNLO/NLO k-factors are applied as a functions of $m_{ZZ}$ and their values vary between 1.0 and 1.2.

Additional NLO electroweak corrections, which depend on the initial state state quark flavour and on kinematics are applied in the mass region $m_{ZZ} > 2m_Z$.

The $gg \rightarrow ZZ$ background MC is generated at the Leading Order (LO) with the generator mcfm 7.0 [37]. Although no exact calculation exists beyond the LO, the soft collinear approximation is able to describe the background cross section and the interference term at NNLO. Further calculations also show that the K-factors are very similar at NLO for the signal and background, suggesting that the same k-factor can be used for the $gg \rightarrow H \rightarrow ZZ$ signal and for the $gg \rightarrow ZZ$ background. The current analysis relies on a NNLO/LO k-factor computed for the signal process as a function of $m_{ZZ}$ and its values vary from about 2.0 to 2.6.

Additional background samples of WZ, Drell-Yan + jets and $t\bar{t}$ are generated using either MadGraph5_aMC@NLO [38] or powheg 2.0, and they are used for the optimization and validation of the data driven methods for the estimation of reducible background. The Drell-Yan + jets sample is also used in lepton-level optimization studies as a source of background leptons.

The list of background samples and their cross sections are shown in Table 5.2.

All signal and background generators are interfaced with PYTHIA 8.2 [39] to simulate multiple parton interactions, underlying event, fragmentation and hadronization effects. Then the generated events are processed through a detailed simulation of the CMS detector based on Geant4 [40] and are reconstructed with the same algorithms that are used for data. The simulated events include overlapping proton-proton interactions (pileup) and are reweighed in order to match the number of interactions per LHC bunch crossing observed in data with the MC simulations.
### Table 5.2: List of background Monte Carlo samples and their corresponding production cross sections at 13 TeV, times relevant branching fractions. In this table $\ell$ means $e$, $\mu$ or $\tau$.

<table>
<thead>
<tr>
<th>Process</th>
<th>$\sigma \times BR$ [pb]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q\bar{q} \rightarrow ZZ \rightarrow 4\ell$</td>
<td>1.256</td>
</tr>
<tr>
<td>$gg \rightarrow ZZ \rightarrow 4e$</td>
<td>0.00159</td>
</tr>
<tr>
<td>$gg \rightarrow ZZ \rightarrow 4\mu$</td>
<td>0.00159</td>
</tr>
<tr>
<td>$gg \rightarrow ZZ \rightarrow 4\tau$</td>
<td>0.00159</td>
</tr>
<tr>
<td>$gg \rightarrow ZZ \rightarrow 2e2\mu$</td>
<td>0.00319</td>
</tr>
<tr>
<td>$gg \rightarrow ZZ \rightarrow 2e2\tau$</td>
<td>0.00319</td>
</tr>
<tr>
<td>$gg \rightarrow ZZ \rightarrow 2\mu2\tau$</td>
<td>0.00319</td>
</tr>
<tr>
<td>$Z \rightarrow \ell\ell + \text{jets}$</td>
<td>6104</td>
</tr>
<tr>
<td>$WZ \rightarrow 3\ell \nu$</td>
<td>4.430</td>
</tr>
<tr>
<td>$t\bar{t}$</td>
<td>815.96</td>
</tr>
</tbody>
</table>

### 5.2 Observables and kinematic discriminants

In general, measuring a physics parameter like a coupling strength, requires to find the observables that are the most sensitive to it. The $H \rightarrow ZZ \rightarrow 4\ell$ channel presents several physics observables, thanks to the possibility to fully reconstruct the final state particles. These observables are built from the event kinematic properties, which include any information that can be computed from the four-momenta of reconstructed objects, like invariant masses and decaying angles. These observables are used either as input to likelihood fits or to categorize events in the analysis.

In this analysis the decay and production kinematics of the Higgs boson are both explored. The full kinematic information for each event is extracted looking at the Higgs boson decay products or the associated particles in its production using kinematic discriminants defined with the so-called matrix element technique. This is a method which builds the discriminant between two processes as the ratio of their matrix elements, if (as is often the case) their phase space factors are identical. These discriminants use a complete set of mass and angular input observables $\hat{\Omega}$ to describe the event kinematics at LO in QCD. The calculations necessary to extract the discriminants are performed with the Matrix Element Likelihood Approach (MELA) package [35], [11], where the JHUGEN [35] and MCFM [37] matrix elements are used for the signal and the background respectively.

#### 5.2.1 Decay observables

The decay part of the kinematic information is used for different purposes, like the separation of the Higgs boson signal from the SM ZZ backgrounds, and the measurement
of its mass, width and spin-parity. In this analysis, two decay observables are defined from the ZZ candidate and are largely used in the analysis: the four-lepton invariant mass $m_{4\ell}$ and a kinematic discriminant $D_{\text{bkg}}^{\text{kin}}$.

The four-lepton invariant mass $m_{4\ell}$ is the most important variable of the analysis. In the $m_{4\ell}$ distribution, the Higgs boson appears as a narrow peak near 125 GeV/c$^2$, on top of a flat background. Its distributions will be shown in the following sections.

The kinematic discriminant $D_{\text{bkg}}^{\text{kin}}$ is defined using the variables that characterize the kinematics of the decay in order to discriminate between the Higgs boson signal and the ZZ background. In fact, the kinematic configuration of a four-lepton system originating from the decay of a Higgs boson is different from that of the four-lepton decay of the irreducible ZZ background.

In the centre-of-mass frame this configuration can be described (up to an arbitrary rotation around the beam axis) by five angles $\theta^*$, $\Phi_1$, $\theta_1$, $\theta_2$ and $\Phi$ (defined in Figure 5.1) and the invariant masses of the two Z candidates, $m_{Z_1}$ and $m_{Z_2}$. These seven variables, collectively referred to as $\vec{\Omega}^{H\rightarrow 4\ell}$, are defined in [35]. The kinematic discriminant sensitive to the the $gg/q\bar{q} \rightarrow 4\ell$ kinematics is calculated as:

$$D_{\text{bkg}}^{\text{kin}} = \left[1 + \frac{\mathcal{P}_{\text{bkg}}^{q\bar{q}}(\vec{\Omega}^{H\rightarrow 4\ell}|m_{4\ell})}{\mathcal{P}_{\text{sig}}^{gg}(\vec{\Omega}^{H\rightarrow 4\ell}|m_{4\ell})}\right]^{-1}$$

(5.2)
Chapter 5. Measurement of Higgs boson properties with 2016 data

5.2 Measurement of Higgs boson properties with 2016 data


de a definition

where $P_{\text{sig}}^{gg}$ is the probability density for an event to be consistent with the signal and $P_{\text{bkg}}^{q\bar{q}}$ is the corresponding probability density for the dominant $q\bar{q} \to ZZ$ background process, both calculated either with the JHUGen or mcfm matrix elements within the MELA framework.

5.2.2 Production observables

Looking at the Feynman diagrams of different Higgs boson production modes (Figure 1.3), it can be noticed that the sub-dominant production modes (VBF, VH, $t\bar{t}H$) present in the final state other particles produced together with the Higgs boson. These associated particles decay according to their nature, producing additional signatures to the required four leptons in each event, like jets or additional leptons not involved in building the best ZZ candidate. Therefore identifying and extracting the Higgs boson production mechanisms considered in this analysis requires to examine also these additional objects. The observables defined from these objects will be used as inputs to the event categorization.

The simplest production observable that can be built is the number of additional objects, like the number of selected jets, b-tagged jets and additional leptons. The number and the type of additional objects varies for each Higgs boson production mode considered. For example VBF and hadronic VH often have one or two reconstructed jets, while b-tagged jets mostly come from $t\bar{t}H$ events. Additional leptons, instead, mostly come from VH leptonic production modes.

The other observables which can be built from additional objects are the production

59
discriminants. They are used to discriminate different production modes according to the additional objects in one event, in particular selected jets, on the basis of their kinematics.

When reconstructed jets are present it is possible to build a discriminant to separate VBF and hadronic VH processes from ggH events produced in association with jets. The production-related kinematics associated to VBF and hadronic VH candidate events is described at LO by five angles illustrated in Figure 5.2 and collectively referred to as $\vec{\Omega}_{H+JJ}$. In events containing at least two selected jets, a discriminant sensitive to the VBF signal topology can be built as:

$$D_{2\text{jet}} = \left[ 1 + \frac{P_{H+JJ}(\vec{\Omega}_{H+JJ}|m_{4\ell})}{P_{VBF}(\vec{\Omega}_{H+JJ}|m_{4\ell})} \right]^{-1}$$

(5.3)

where the $P_{H+JJ}$ and $P_{VBF}$ probabilities are obtained from the JHUGen matrix elements. This discriminant separates VBF signal from ggH + 2jets events.

In VBF events it often happens that one of the two jets is out of the detector acceptance or is not reconstructed or fails the selection requirements. A VBF signal probability can be built in events containing exactly one selected jet integrating the $P_{VBF}$ over the pseudorapidity of the unobserved jet, constraining the transverse momentum of the $4\ell + 2$ jets system to zero. The probability for ggH + 1jet events is computed with the JHUGen matrix elements, allowing the definition of the discriminant:

$$D_{1\text{jet}} = \left[ 1 + \frac{P_{H+JJ}(\vec{\Omega}_{H+JJ}|m_{4\ell})}{\int d\eta_{jj} P_{VBF}(\vec{\Omega}_{H+JJ}|m_{4\ell})} \right]^{-1}$$

(5.4)

which separates VBF signal with 1 jet from ggH + 1jets events.

Two other discriminants are built for events with at least two selected jets in order to separate VH hadronic processes from ggH + 2jets events, as:

$$D_{WH} = \left[ 1 + \frac{P_{H+JJ}(\vec{\Omega}_{H+JJ}|m_{4\ell})}{P_{WH}(\vec{\Omega}_{H+JJ}|m_{4\ell})} \right]^{-1}$$

(5.5)

$$D_{ZH} = \left[ 1 + \frac{P_{H+JJ}(\vec{\Omega}_{H+JJ}|m_{4\ell})}{P_{ZH}(\vec{\Omega}_{H+JJ}|m_{4\ell})} \right]^{-1}$$

(5.6)

where the $P_{WH}$ and $P_{ZH}$ probabilities are obtained from the JHUGen matrix elements.

Matrix element probabilities for VH leptonic processes have also been studied but they are not used in the analysis because they were found to add little information to the tagging provided by the simple presence of additional reconstructed leptons.

By construction, all discriminants defined have values bounded between 0 and 1. They are used to enhance the purity of event categories, described in the following section.
5.3 Categorizing events

In order to improve the sensitivity to the different Higgs boson production mechanisms, the selected events are classified into mutually exclusive categories, exploiting the additional signatures of events, considering the observables defined in the previous section. Seven categories are defined according to the jet multiplicity, the number of b-tagged jets, the number of additional leptons (reconstructed leptons which do not form the ZZ candidate) and requirements on the kinematic discriminants, using the following criteria applied in the exact presented order (i.e., an event is considered for the following category only if it does not satisfy the requirements of the previous category):

- **VBF-2jet-tagged category** requires exactly four leptons. In addition, there must be either two or three jets of which at most one is b-tagged, or at least four jets and no b-tagged jets. \( D_{2\text{jet}} > 0.5 \) is then required.
- **VH-hadronic-tagged category** requires exactly four leptons. In addition, there must be two or three jets, or at least four jets and no b-tagged jets. Then \( D_{\text{VH}} \equiv \max(D_{\text{WH}}, D_{\text{ZH}}) > 0.5 \) is required.
- **VH-leptonic-tagged category** requires no more than three jets and no b-tagged jets in the event, and exactly one additional lepton or one additional pair of OS, same flavour leptons. This category also includes events with no jets and at least one additional lepton.
- **\( t\bar{t}H \)-tagged category** requires at least four jets of which at least one is b tagged, or at least one additional lepton.
- **VH-\( E_T^{\text{miss}} \)-tagged category** requires exactly four leptons, no more than one jet and \( E_T^{\text{miss}} > 100 \text{ GeV} \).
- **VBF-1jet-tagged category** requires exactly four leptons, exactly one jet and \( D_{1\text{jet}} > 0.5 \).
- **Untagged category** consists of the remaining selected events.

In each of the defined categories the requirements and the cuts on the physics observables are optimized in order to maximize the purity of the tagged production mode, without penalizing the number of events in each category. Thanks to the purity of the categories, it is possible to enhance the sensitivity of the analysis to the Higgs boson production mechanisms and the amount of event in each category allows the significance of the signal to be preserved.

Each category is defined in order to select a certain type of events coming from a particular production mode. The gluon fusion production mode produces events with no additional signatures in the final state, therefore ggH events are collected in the **Untagged category**, which contains all the events that do not satisfy the other categories.
Chapter 5. Measurement of Higgs boson properties with 2016 data

Figure 5.3: Relative signal purity in the seven event categories in terms of the five main production mechanisms of the Higgs boson in the in the 118 < m_{4\ell} < 130 GeV/c^2 mass window. The WH, ZH, and t\bar{t}H processes are split according to the decay of the associated particles, where X denotes anything other than an electron or muon. Numbers indicate the total expected signal event yields in each category [1].

criteria. The **VBF-2jet-tagged category** is defined in order to select events coming from the VBF production mode where the majority of the jets, produced by the hadronization of the two quarks which are present in the final state, are selected by the detectors. The **VBF-1jet-tagged category**, instead, collects the events produced with VBF but where only one jet is reconstructed inside the detector acceptance. The **VH-hadronic-tagged category** is intended to select events coming from the VH production mode where the associated vector boson decays hadronically, therefore the presence of additional jets is required. Furthermore the request on the \(D_{VH}\) discriminant is necessary in order to exclude events coming from the VBF production mode, where similarly the presence of two jets is expected. The **VH-leptonic-tagged category** selects events from VH mechanism where the vector boson decays leptonically and thus additional leptons are required. The **ttH-tagged category** is defined to collect events in which the Higgs boson is produced with a t\bar{t} quark pair, and therefore the presence of a b-tagged jet is required in order to account the top quark decay into a W boson and a b quark. The **VH-\(E_T^{miss}\)-tagged category** selects events coming from VH mechanisms where the associated boson decays in neutrinos, or hadronically in the case the jets are not reconstructed inside the detector acceptance and therefore a certain amount of missing energy is required by the categorization in order to account these kind of decays.
Chapter 5. Measurement of Higgs boson properties with 2016 data

The order in which the categories and therefore the order in which the selection cuts are applied is optimized to obtain the maximum value of signal significance for each category. Figure 5.3 shows the relative signal purity of the seven event categories for the various Higgs boson production processes.

The expected number of events in some of these categories is very low, in particular in the VH-$E_T^{miss}$-tagged category. The definition of these categories has been made for the future, in the perspective of the expected integrated luminosity at the end of LHC Run II (about 150 fb$^{-1}$). That amount of data will allow all categories to be populated in a significant way.

5.4 Inputs for the statistical model

Once the events have been selected and the ZZ candidates have been built as described in Chapter 4 and they have been divided into categories, as described in Section 5.3, a statistical analysis has to be performed in order to extract results from selected data. The statistical analysis is performed with the statistical model described in Chapter 4. In order to perform the analysis with the statistical model it is necessary to first provide the inputs of the model: expected signal and background yields, shapes and systematic uncertainties. The estimation of the model inputs is presented in the following sections.

Since in the analysis the selected events are divided into categories in order to improve the sensitivity of the analysis to the Higgs boson production modes, the inputs of the statistical model have to be provided separately for each category. Moreover the defined categories are not completely pure: the events contained inside a given category may come from different production modes. For this reason it is necessary to know exactly the composition of each category in terms of production modes.

Furthermore the estimation of model inputs is performed separately for the three possible final states ($4\mu$, $4e$, and $2e2\mu$) since there are differences in background sources, resolution and systematic uncertainties sources in these three channels.

5.4.1 Event yields

For the current analysis, the number of expected background and signal events are estimated from MC simulations both for signal and irreducible background samples, while for the reducible background $Z+X$ are estimated with a data-driven technique.

Signal expected yields

Signal expected yields are the normalization of the Higgs boson signal in the peak region. They are computed using MC samples for five Higgs boson mass points (120, 124, 125, 126, and 130 GeV/$c^2$) and parametrized as a function of $m_H$ with polynomial fits in a
Chapter 5. Measurement of Higgs boson properties with 2016 data

Figure 5.4: Expected signal yield as a function of the Higgs boson mass $m_H$ for an integrated luminosity of 35.9 fb$^{-1}$ in the $105 < m_{4\ell} < 140$ GeV/$c^2$ window for the $ggH$ production mode in the Untagged category (left), and for the VBF production mode in the VBF2jTagged category (right).

four-lepton invariant mass $m_{4\ell}$ window around the Higgs boson peak $[105, 140]$ GeV/$c^2$. The fits are performed separately for the seven considered production modes (ggH, VBF, WH hadronic, WH leptonic, ZH hadronic, ZH leptonic, and t\bar{t}H), the three final states and the seven categories. Some examples of the yields parametrizations are shown in Figure 5.4 for the ggH production mode in the Untagged category (left frame) and for the VBF production mechanism in the VBF2jTagged category (right frame).

A continuous parametrization of the signal predictions as a function of the Higgs boson mass is necessary, for example, in order to measure the signal strength parameter for a specific value of $m_H$.

**Background estimation**

Two types of background are considered in the analysis:

- **Irreducible background**, consisting in processes which present exactly the same final state as the signal.

- **Reducible background**, consisting in events which pass the selection because of mis-reconstructed or mis-identified objects, due to the non perfect performance of the detector and of the reconstruction techniques.

**Irreducible background** Irreducible background for the $H \rightarrow ZZ \rightarrow 4\ell$ signal comprehends processes which present two pairs of opposite-sign same-flavour leptons in their final state.
The expected yields of the two irreducible background contributions ($q\bar{q} \rightarrow ZZ$ and $gg \rightarrow ZZ$) to the signal region is estimated using MC simulations. These contributions have to be corrected for higher order corrections that are not taken into account in the MC generation. This is done by multiplying the expected event yields by the k-factors, which contain the contributions for higher order corrections. After these corrections are applied, the expected yields are computed in the mass window $105 < m_{4\ell} < 140 \text{ GeV}/c^2$, separately for every final state and event category.

**Reducible background** The reducible background for the $H \rightarrow ZZ \rightarrow 4\ell$ analysis is denoted as $Z+X$ since, because of the selection requirements, it appears as a reconstructed $Z$ made by real prompt leptons, accompanied by a second lepton pair including one or two non-prompt leptons. It can be due to leptons from heavy-flavour hadron decays or light mesons mis-identified as leptons.

The contribution of the $Z+X$ background to the signal region can not be accurately estimated using MC because it involves several physics processes which can not be simulated with sufficient statistics to correctly populate the four-lepton signal region. Therefore it is estimated using two independent data-driven methods using data events collected in the control regions, which are regions of events with similar properties to the signal, but depleted in the actual signal content and enriched in the kind of background events that must be estimated.

The control regions are defined by a di-lepton pair satisfying all the requirements of a $Z_1$ candidate and two additional leptons of opposite sign (OS) or same-sign (SS), satisfying certain relaxed identification requirements when compared to those used in the analysis. In the case of OS leptons the signal events are excluded by inverting some selection requirements, while in the SS case signal events are excluded by the same-sign leptons request itself. The four selected leptons are then required to pass the ZZ candidate selection.

The background yield in the signal region is estimated by weighting the control region events by the lepton misidentification probability (or fake rate) $f$, defined as the probability for a lepton passing the relaxed identification requirements used in the given control region to actually pass the full identification requirements. The electron and muon fake rates, $f_e$ and $f_\mu$, are determined from data, separately for the SS and OS methods, using a different control region defined by a $Z_1$ candidate and exactly one additional lepton passing the relaxed selection. The $Z_1$ candidate consists of a pair of leptons, each of which passes the selection requirements used in the analysis. For the OS method, the mass of the $Z_1$ candidate is required to satisfy $|m_{Z_1} - m_Z| < 7 \text{ GeV}/c^2$ to reduce the contribution of photon conversions, which is estimated separately. In the SS method, the contribution from photon conversions is estimated by determining an average misidentification rate. Furthermore the $E_T^{\text{miss}}$ is required to be less than $25 \text{ GeV}/c^2$ to suppress contamination from WZ and tt processes.

The **OS method** utilizes events from the control region with a $Z_1$ candidate and two
additional OS leptons of the same flavour. The expected yield in the signal region is obtained from two categories of events.

The first category is composed of events with two leptons that pass (P) the tight lepton identification requirements and two leptons that pass the loose identification but fail (F) the tight identification, and is denoted as the 2P2F region. Backgrounds, which intrinsically have only two prompt leptons, such as Z + jets, t\(\bar{t}\)+jets, are estimated with this control region. To obtain the expected yield in the signal region, each event in the 2P2F region is weighted by a factor \(\frac{f_i^3}{1 - f_i^3} \frac{f_i^4}{1 - f_i^4}\), where \(f_i^3\) and \(f_i^4\) are the misidentification rates for the third and fourth lepton, respectively, that do not pass the tight identification.

The second category consists of events where exactly one of the two additional leptons passes the analysis selection, and is referred to as the 3P1F region. Backgrounds with three prompt leptons, such as WZ + jets and Z\(\gamma\) + jets with photon converting to e\(^+\)e\(^-\), are estimated using this region. To obtain the expected yield in the signal region, each event in the 3P1F region is weighted by a factor \(\frac{f_j^4}{1 - f_j^4}\), where \(f_j^4\) is the misidentification rate for the lepton that does not pass the analysis selection. The contribution from ZZ events to the 3P1F region (\(N_{ZZ}^{3P1F}\)) which arises from events where a prompt lepton fails the identification requirements, is estimated from simulation and scaled with a factor \(w_{ZZ}\) appropriate to the integrated luminosity of the analysed dataset. The contamination of 2P2F-type processes in the 3P1F region is subtracted from the total background estimate to avoid double counting.

The total reducible background estimate in the signal region coming from the two categories 2P2F and 3P1F without double counting can be written as:

\[
N_{\text{reducible}}^{\text{SR}} = N_{3P1F} \sum_j \frac{f_j^4}{1 - f_j^4} - w_{ZZ} N_{3P1F} \sum_j \frac{f_j^4}{1 - f_j^4} - \sum_i \frac{f_i^3}{1 - f_i^3} \frac{f_i^4}{1 - f_i^4}
\]

where \(N_{3P1F}\) and \(N_{2P2F}\) are the number of events in the 3P1F and 2P2F regions, respectively.

The SS method utilizes a control region, referred to as 2P2L\(_{SS}\) region, which consists of events with a Z\(_1\) candidate two additional SS leptons of same flavour. These two additional leptons are required to pass the loose selection requirements for leptons. The contribution of photon conversions to the electron misidentification probability \(f\) is estimated using its dependence on the fraction of loose electrons in the sample with tracks having one missing hit in the pixel detector.

The expected number of reducible background events in the signal region can then be written as:

\[
N_{\text{reducible}}^{\text{SR}} = r_{\text{OS/SS}} \sum_i \frac{N_{3P2LSS}}{f_3 f_4}
\]

where the ratio \(r_{\text{OS/SS}}\) between the number of events in the 2P2L\(_{OS}\) and 2P2L\(_{SS}\) control
regions is obtained from simulation. The 2P2L_{OS} region is defined analogously to the 2P2L_{SS} region but with an OS requirement for the additional pair of loose leptons.

The predicted yield in the signal region of the reducible background from the two methods are in agreement within their statistical uncertainties, and since they are mutually independent, the results of the two methods are combined. The final estimate is obtained by weighting the individual mean values of both methods according to their corresponding variances.

5.4.2 Mass lineshape

For the current analysis the signal and backgrounds are modelled with dedicated PDFs in order to perform the shape analysis.

Signal lineshape

The signal lineshapes are modelled separately for the seven considered Higgs boson production modes (ggH, VBF, WH hadronic, WH leptonic, ZH hadronic, ZH leptonic, and t\bar{t}H), the three final states and the seven categories. The analysis exploits the $105 < m_{4\ell} < 140 \text{ GeV}/c^2$ invariant mass $m_{4\ell}$ window and the lineshape is built for the $118 < m_{4\ell} < 130 \text{ GeV}/c^2$ mass range, using five Higgs boson mass points (120, 124, 125, 126, and 130 \text{ GeV}/c^2).

A SM Higgs boson resonance of mass around 125 GeV/c$^2$ can be described in the narrow-width approximation, therefore the theoretical signal lineshape can be described as a relativistic Breit-Wigner function. This theoretical lineshape has to be convolved with an empirical function that accounts for the effects of experimental resolution, such as tails due to bremsstrahlung, final state radiation, and energy leakage in the ECAL. In this analysis a Double Crystal Ball (DCB) function, which accounts for the Gaussian resolution of the core of the invariant mass distribution and includes the two asymmetric tails described by power laws. Since the contribution of the intrinsic peak width is negligible if compared to the experimental distribution, the PDF used to describe the signal shapes is defined by a DCB function alone:

$$\text{DCB}(\zeta) = N \cdot \begin{cases} A \cdot (B + |\zeta|)^{-n_L}, & \text{for } \zeta < \alpha_L \\ \exp(-\zeta^2/2), & \text{for } \alpha_L \leq \zeta \leq \alpha_R \\ A \cdot (B + |\zeta|)^{-n_R}, & \text{for } \zeta > \alpha_R \end{cases} \quad (5.9)$$

where $\zeta = (m_{4\ell} - m_H - \Delta m_H)/\sigma_m$. This function has six independent parameters: the width $\sigma_m$ of the Gaussian core, the systematic mass shift $\Delta m_H$ of the reconstructed peak, the two parameters $n_L$ and $n_R$ which control the prominence of each tail, and the two parameters $\alpha_L$ and $\alpha_R$ which define the position of the boundary between the core and the tails. The parameters $A$ and $B$ are defined by requiring the continuity of the function and its first derivative, and the parameter $N$ is the normalizing constant.
Figure 5.5: MC simulated $m_{4\ell}$ distributions (markers) and the parametrized lineshapes (lines) for the the ggH process generated with $m_H = 125\text{ GeV}/c^2$, in the Untagged category, for the three final states: 4$\mu$ (top left), 4$e$ (top right) and 2e2$\mu$ (bottom).

The dependency of each of the six parameters $p_i$ on $m_H$ is parametrized for each final state and event category as:

$$p_i(m_H) = p^i_{\text{DCB0}} + p^i_{\text{DCB1}} \times (m_H - 125\text{ GeV}/c^2)$$  \hspace{1cm} (5.10)

where the $p^i_{\text{DCB0}}$ are the parameters of the DCB obtained from shapes in the ggH production mode in the Untagged category for $m_H = 125\text{ GeV}/c^2$, while $p^i_{\text{DCB1}}$ are the parameters obtained form a simultaneous fit for all the other $m_H$ points.

Figure 5.5 shows the signal lineshapes for the ggH production mode in the Untagged category for the three different final states, computed for the Higgs boson mass point $m_H = 125\text{ GeV}/c^2$.

As for the ggH process, the signal lineshape of the VBF process is described by a DCB function only. For the other considered processes (WH, ZH, $t\bar{t}H$), instead, the Higgs boson peak is accompanied by a non-resonant component due to additional leptons.
Chapter 5. Measurement of Higgs boson properties with 2016 data

Figure 5.6: MC simulated $m_{4\ell}$ distributions (markers) and parametrized lineshapes (lines) for the the ZH process generated with $m_H = 125 \text{ GeV}/c^2$, in the Untagged category, for the three final states: 4\(\mu\) (top left), 4\(e\) (top right) and 2\(e\)2\(\mu\) (bottom). The pink line represents the Landau component, enlarged in the picture for clarity, and the light green shape represents the DCB contribution.

coming from the decay of the associate particles produced with the Higgs boson in these production modes. The non resonant tail appears when one of these additional leptons is taken to built the four-lepton candidate in place of one of the Z boson decay leptons form $H \rightarrow ZZ$ which may be out of the detector acceptance. Therefore, this non resonant component is described with an additional Landau function, requiring two more parameters to the total PDF. The relative normalization of both components is adjusted for each category. Figure 5.6 presents the signal lineshapes for the ZH process showing the two components of the DCB and the Landau functions.

Background lineshape

The irreducible background mass lineshapes are obtained fitting the MC histograms with empirical functions, Bernstein polynomials of degree 3, separately for the two irreducible backgrounds $q\bar{q} \rightarrow ZZ$ and $gg \rightarrow ZZ$, and for every final state and event category in the mass range $105 < m_{4\ell} < 140 \text{ GeV}/c^2$. The MC events are weighed with higher order corrections (k-factors).

The reducible background shape is obtained from the prediction of the OS method (described in section 5.4.1) and fitting the event distributions with empirical functional forms built from Landau and exponential functions.
5.4.3 Systematic uncertainties

Each source of systematic uncertainty considered in the $H \to ZZ \to 4\ell$ analysis is assigned to a nuisance parameter $\theta_i$ which enters in the statistical model as explained in section 4.1.2. It can affect one or multiple processes, in different ways according to the final state or the event category. The exact way and magnitude in which a nuisance parameter affects a certain process is determined and quantified from specific studies. The considered systematics include all the sources of uncertainties which make the prediction of signal and background yields and shapes imperfect, biasing the comparison with collected data. They comprehend two kind of uncertainties:

- **Theoretical**: related to cross section computations and theoretical hypotheses at the basis of MC generators;
- **Experimental**: related to the detector response, like efficiencies and calibration, and to the estimation of the $Z+X$ background from data.

The uncertainties considered in the analysis can affect both expected yields and mass lineshapes. Therefore there are two classes of uncertainties considered:

- **Normalization uncertainties**, which affect the expected yields of some processes. Their associated nuisance parameters scale the expected yields of a certain value, and their PDFs follow a log-normal distribution.
- **Shape uncertainties**, which alter the mass lineshapes of some processes. This is modelled by considering a family of alternative shapes, which is governed by one single nuisance parameter that itself follows a Gaussian distribution.

All the systematics presented in the following sections are treated as uncorrelated one to another, and unless otherwise specified, they are normalization uncertainties.

**Theoretical uncertainties**

These are systematic uncertainties due to theoretical computations.

- A **QCD uncertainty** is applied to every signal and irreducible background sample in an uncorrelated way. The renormalization and factorization QCD scale variation has impact on the global process normalization (for the 3-10%) and also causes some migrations of events between different categories.
- The QCD uncertainty for the ggH process contains also the effect of the **modelling of hadronization**, which has been studied only for this process based on dedicated generator-level samples. The resulting uncertainty is found to be 6% and it is added linearly to the ggH QCD uncertainty.
Chapter 5. Measurement of Higgs boson properties with 2016 data

- The uncertainty coming from the choice of the set of parton distribution functions at the generation level is determined independently for different sets of processes, grouped by initial state, like $gg \to ZZ$ and $gg \to H \to ZZ$. This uncertainty is computer inclusively and also for each category.

- To the q̅q → ZZ background prediction, systematic uncertainties on the electroweak corrections are applied as mass-dependent uncertainties. They are normalization uncertainties which account for event migration from high-mass and their value varies between 1 to 8% according to the event category.

- An additional normalization uncertainty of 10% is applied to the $gg \to ZZ$ yield, in order to account that the k-factor applied to this process was actually computed for the signal process $gg \to H \to ZZ$. The contribution of this uncertainty is added to the QCD systematic for this background process.

- A systematic uncertainty of 2% on the branching ratio of the $H \to ZZ \to 4\ell$ channel is applied to the yields of all signal processes.

Experimental uncertainties

Uncertainties coming from experimental features can be due to several sources.

- The measurement of the integrated luminosity of the data sample is affected by a normalization uncertainty. This is defined for all the CMS analyses for a given data taking period and for the 2016 analysis its contribution is 2.6%, applied to all processes.

- The imprecise knowledge of the jet energy scale affects the number of jets considered in a certain event. This variable is used in the event categorization, therefore a variation in the jet energy scale cause migration of events from one category to another. This systematic uncertainty does not affect the overall normalization or the lineshapes, and it is fully correlated for different processes and anti-correlated between categories. Its contribution varies from 2 to 18% according to the process and the event category.

- A systematic uncertainty is assigned to the b-tagging efficiency, which affects the number of b-tagged jets per event. This causes event migrations between categories and can alter the expected yields by 2-8% in those categories defined according to the b-jet number. The impact on the other categories is lower. As the previous systematic, also this one is fully correlated for different processes and anti-correlated between categories.

- Efficiency of trigger and lepton reconstruction and selection uncertainties are covered by one flat value for each final state. This value is obtained by propagating the per-lepton uncertainties defined as the quadrature sum of the per-lepton uncertainty on trigger efficiency and that on scale factors for lepton reconstruction.
Chapter 5. Measurement of Higgs boson properties with 2016 data

and selection efficiencies (cf. Eq. 3.7). The propagation is performed by assuming full correlation between different per-event leptons of the same flavour, and assuming leptons of different flavour as uncorrelated.

- The uncertainty on the lepton energy scale comes from the lepton energy calibration procedure. This is a shape uncertainty on $m_{4\ell}$ which describes the Gaussian variation of the mean of the mass lineshape. This uncertainty is relevant only for the signal samples and varies between 0.3% for electrons and 0.04% for muons.

- Another shape uncertainty affects the 4\ell mass resolution ($\sigma_m$). This uncertainty describes the Gaussian variation of the sigma of the mass lineshape and if differs for the three final states: 20% for 4e, 10% for 4\mu while for 2e2\mu it is computed as the mean of the 4e and 4\mu values.

- The reducible background estimation is affected by uncertainties which are due to the difference in the composition of the Z + \ell control region used to compute the fake rates and the control regions where the fake rates are applied. The uncertainties are treated as uncorrelated between final states and different fixed values are assigned for each final state: 32% for 4e, 35% for 4\mu and 34% for 2e2\mu. The Z + X background is also affected by a shape uncertainty which account the fact that the q\bar{q} \rightarrow ZZ background has a similar shape. The uncertainty on the Z + X shape is defined as the difference between these two shapes.

5.5 Results of H → ZZ → 4\ell 2016 analysis

This section presents the outcome of event selection and the distributions of the main kinematic variables in data and MC, which are used as inputs of the statistical analysis performed to extract the measurement of Higgs boson properties, shown at the end of this section.

5.5.1 Event yields

The number of events observed in the 2016 data sample corresponding to an integrated luminosity of 35.9 fb$^{-1}$ is presented in Table 5.3. The events are selected according to the criteria presented in Chapter 3. They are reported in the table together with the expected background and signal events both for the full 4\ell sample and for each final state, in the full mass range $m_{4\ell} > 70$ GeV/$c^2$. The q\bar{q} → ZZ and gg → ZZ background yields are estimated from MC simulation, while the Z+X yield is computed with data-driven methods reported in Section 5.4.1.

Table 5.4 instead, shows the expected and observed yields, focusing this time on the 118 < $m_{4\ell}$ < 130 GeV/$c^2$ mass window around the mass of the Higgs boson. To obtain the Z+X expected yield in this mass range, the prediction of the full signal region is scaled according to the relevant fraction of the integral of the full $m_{4\ell}$ analytical shape.
Chapter 5. Measurement of Higgs boson properties with 2016 data

<table>
<thead>
<tr>
<th>Channel</th>
<th>4e</th>
<th>4μ</th>
<th>2e2μ</th>
<th>4ℓ</th>
</tr>
</thead>
<tbody>
<tr>
<td>qq → ZZ</td>
<td>193$_{+139}^{−20}$</td>
<td>360$_{+25}^{−27}$</td>
<td>471$_{+33}^{−36}$</td>
<td>1024$_{+102}^{−70}$</td>
</tr>
<tr>
<td>gg → ZZ</td>
<td>41$_{−6}$</td>
<td>69.0$_{−3.0}$</td>
<td>102$_{−13}$</td>
<td>212$_{−27}$</td>
</tr>
<tr>
<td>Z + X</td>
<td>21.1$_{−10.4}^{+8.3}$</td>
<td>34$_{−13}^{+14}$</td>
<td>60$_{−25}^{+27}$</td>
<td>115$_{−30}$</td>
</tr>
<tr>
<td>Sum of backgrounds</td>
<td>255$_{−25}^{+24}$</td>
<td>463$_{−34}^{+32}$</td>
<td>633$_{−46}^{+44}$</td>
<td>1351$_{−91}^{+86}$</td>
</tr>
<tr>
<td>Signal (m_H = 125 GeV/c^2)</td>
<td>12.0$_{−1.4}^{+1.5}$</td>
<td>23.6$_{−2.1}^{+2.4}$</td>
<td>30.0$_{−2.6}^{+2.3}$</td>
<td>65.7$_{−5.6}^{+5.4}$</td>
</tr>
<tr>
<td>Total expected</td>
<td>267$_{−26}^{+23}$</td>
<td>487$_{−35}^{+33}$</td>
<td>663$_{−47}^{+46}$</td>
<td>1417$_{−94}^{+89}$</td>
</tr>
<tr>
<td>Observed</td>
<td>293</td>
<td>505</td>
<td>681</td>
<td>1479</td>
</tr>
</tbody>
</table>

Table 5.3: Number of expected background and signal events and of observed events after the full selection, for each final state, for the full mass range m_{4ℓ} > 70 GeV/c^2, for an integrated luminosity of 35.9 fb$^{-1}$. The signal and ZZ backgrounds are estimated from simulation, while the Z+X event yield is estimated from data. Uncertainties include statistical and systematic sources [1].

Similarly, Table 5.5 reports the number of observed and expected events for each of the seven event categories in the 118 < m_{4ℓ} < 130 GeV/c^2 mass range. The quoted uncertainties include systematic sources and, in the case of Z+X, the systematic uncertainties deriving from the statistical uncertainty in the control regions used for estimating this background.

5.5.2 Event distributions

The reconstructed four-lepton invariant mass distribution is shown in Figure 5.7 for the full dataset corresponding to an integrated luminosity of 35.9 fb$^{-1}$, both for the full mass range and in the region around the Higgs boson peak. The Higgs boson appears as a narrow peak on a flat background distribution. The observed data are compared with the expectations for signal and background processes. Points with error bars represent the data while stacked histograms represent expected signal and background distributions. The SM Higgs boson signal and the ZZ backgrounds are normalized to the SM expectation, while the Z+X background is normalized to the estimation from data. The error bars on the data points correspond to 68% confidence intervals. The observed distribution agrees with the expectation within the statistical uncertainties over the whole invariant mass spectrum.

In Figure 5.8 the four-lepton invariant mass distributions split by event category for the low-mass range are presented. As in the inclusive plots the SM Higgs boson signal and the ZZ backgrounds are normalized to the SM expectation and the Z+X background is normalized to the estimation from data, but the signal sample (Higgs boson signal with m_H = 125 GeV/c^2) is divided in two contributions: one is the production mode which is targeted by the considered category, and the other which includes the remaining production modes among which the gluon fusion mechanism dominates. Also in these plots the agreement between data and MC is good, considering the statistical uncertainties.
Chapter 5. Measurement of Higgs boson properties with 2016 data

Table 5.4: Number of expected background and signal events and of observed events after the full selection, for each final state, for the 118 < $m_{4\ell}$ < 130 GeV/$c^2$ mass range, for an integrated luminosity of 35.9 fb$^{-1}$. The signal and ZZ background expected yields are estimated from simulation, while the Z+X event yield is estimated from data. Uncertainties include statistical and systematic sources [1].

<table>
<thead>
<tr>
<th>Channel</th>
<th>$4\ell$</th>
<th>$4\mu$</th>
<th>$2e2\mu$</th>
<th>$4\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>qq → ZZ</td>
<td>$3.23^{+0.34}_{-0.34}$</td>
<td>$8.56^{+0.09}_{-0.09}$</td>
<td>$10.22^{+0.71}_{-0.77}$</td>
<td>$22.01^{+1.37}_{-1.62}$</td>
</tr>
<tr>
<td>gg → ZZ</td>
<td>$0.40^{+0.06}_{-0.06}$</td>
<td>$0.88^{+0.12}_{-0.11}$</td>
<td>$0.81^{+0.11}_{-0.10}$</td>
<td>$2.09^{+0.25}_{-0.27}$</td>
</tr>
<tr>
<td>Z + X</td>
<td>$1.93^{+0.78}_{-0.75}$</td>
<td>$4.71^{+1.99}_{-1.81}$</td>
<td>$6.88^{+3.41}_{-2.87}$</td>
<td>$13.52^{+5.74}_{-3.53}$</td>
</tr>
<tr>
<td>Sum of backgrounds</td>
<td>$5.57^{+0.83}_{-0.81}$</td>
<td>$14.15^{+1.69}_{-1.90}$</td>
<td>$18.91^{+2.35}_{-2.45}$</td>
<td>$37.63^{+3.44}_{-3.45}$</td>
</tr>
<tr>
<td>Signal ($m_H = 125$ GeV/$c^2$)</td>
<td>$10.18^{+1.14}_{-1.15}$</td>
<td>$21.62^{+1.91}_{-1.90}$</td>
<td>$26.52^{+2.24}_{-2.32}$</td>
<td>$58.32^{+5.05}_{-5.02}$</td>
</tr>
<tr>
<td>Total expected</td>
<td>$15.75^{+1.00}_{-1.02}$</td>
<td>$35.77^{+2.92}_{-2.95}$</td>
<td>$44.43^{+3.66}_{-3.64}$</td>
<td>$95.95^{+6.74}_{-6.74}$</td>
</tr>
<tr>
<td>Observed</td>
<td>19</td>
<td>34</td>
<td>41</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 5.5: Number of expected background and signal events and of observed events after the full selection, for each event category, for the mass range 118 < $m_{4\ell}$ < 130 GeV/$c^2$. The yields are presented for the different production modes. The signal and ZZ backgrounds expected yields are estimated from simulation, while the Z+X event yield is estimated from data. [1]

<table>
<thead>
<tr>
<th>Channel</th>
<th>Untagged</th>
<th>VBF-1j</th>
<th>VBF-2j</th>
<th>VH-had</th>
<th>VH-lep</th>
<th>VH-E$^{miss}$</th>
<th>ttH</th>
<th>Inclusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>qq → ZZ</td>
<td>19.18</td>
<td>2.00</td>
<td>0.25</td>
<td>0.30</td>
<td>0.27</td>
<td>0.01</td>
<td>0.01</td>
<td>22.01</td>
</tr>
<tr>
<td>gg → ZZ</td>
<td>1.67</td>
<td>0.31</td>
<td>0.05</td>
<td>0.02</td>
<td>0.04</td>
<td>0.01</td>
<td>&lt;0.0</td>
<td>2.09</td>
</tr>
<tr>
<td>Z + X</td>
<td>10.79</td>
<td>0.88</td>
<td>0.78</td>
<td>0.31</td>
<td>0.17</td>
<td>0.30</td>
<td>0.27</td>
<td>13.52</td>
</tr>
<tr>
<td>ggH</td>
<td>31.64</td>
<td>3.18</td>
<td>1.08</td>
<td>0.63</td>
<td>0.49</td>
<td>0.32</td>
<td>0.28</td>
<td>37.62</td>
</tr>
<tr>
<td>VBF</td>
<td>1.08</td>
<td>1.14</td>
<td>2.09</td>
<td>0.09</td>
<td>0.02</td>
<td>&lt;0.01</td>
<td>&lt;0.0</td>
<td>4.44</td>
</tr>
<tr>
<td>WH</td>
<td>0.33</td>
<td>0.14</td>
<td>0.05</td>
<td>0.30</td>
<td>0.21</td>
<td>0.03</td>
<td>0.02</td>
<td>1.18</td>
</tr>
<tr>
<td>ZH</td>
<td>0.41</td>
<td>0.11</td>
<td>0.04</td>
<td>0.24</td>
<td>0.04</td>
<td>0.07</td>
<td>0.02</td>
<td>0.93</td>
</tr>
<tr>
<td>tth</td>
<td>0.08</td>
<td>&lt;0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>&lt;0.01</td>
<td>0.01</td>
<td>0.50</td>
</tr>
<tr>
<td>Signal</td>
<td>40.77</td>
<td>9.09</td>
<td>4.24</td>
<td>2.08</td>
<td>0.38</td>
<td>0.11</td>
<td>0.51</td>
<td>57.79</td>
</tr>
<tr>
<td>Total expected</td>
<td>72.41</td>
<td>12.88</td>
<td>5.32</td>
<td>2.71</td>
<td>0.86</td>
<td>0.43</td>
<td>0.79</td>
<td>95.41</td>
</tr>
<tr>
<td>Observed</td>
<td>73</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>94</td>
</tr>
</tbody>
</table>

74
Figure 5.7: Distribution of the reconstructed four-lepton invariant mass $m_{4\ell}$ in the full mass range (top frame) and the low-mass range (bottom frame). The SM Higgs boson signal and the ZZ backgrounds are normalized to the SM expectation, while the Z+X background is normalized to the estimation from data [1].
Figure 5.8: Distribution of the reconstructed four-lepton invariant mass in the seven event categories for the low-mass range. Untagged category (top left), VBF-1jet-tagged category (centre left), VH-hadronic-tagged category (centre right), VH-leptonic-tagged category (bottom left), VH-$E_T^{miss}$-tagged category (bottom middle), $t\bar{t}H$-tagged category (bottom right). For the categories other than the untagged category, the SM Higgs boson signal is separated into two components: the production mode that is targeted by the specific category, and other production modes, where the gluon fusion dominates [1].
Chapter 5. Measurement of Higgs boson properties with 2016 data

Figure 5.9: Distribution of $D_{\text{kin}}^{\text{bkg}}$ vs. $m_4\ell$ in the mass region $100 < m_4\ell < 170 \text{ GeV}/c^2$. The grey scale represents the expected total number of ZZ background and SM Higgs boson signal events for $m_H = 125 \text{ GeV}/c^2$. The points show the data and the horizontal bars represent the four-lepton mass uncertainties. Different marker colours and styles are used to denote final state and the categorization of the events, respectively [1].

In Figure 5.9 the correlation between the kinematic discriminant $D_{\text{kin}}^{\text{bkg}}$ with the four-lepton invariant mass $m_4\ell$ is shown. The grey scale represents the expected combined relative density of the ZZ background and the Higgs boson signal. The points show the data and the measured four-lepton mass uncertainties as horizontal bars. Different marker colours and styles are used to denote the final state and the categorization of the events, respectively. This distribution shows that the two observed events around 125 GeV/$c^2$ in the VH-$E_T^{\text{miss}}$ and $t\bar{t}H$ categories have low values of $D_{\text{kin}}^{\text{bkg}}$, which means that these events are more compatible with the background than the signal hypothesis.

5.5.3 Significance and signal strength modifiers

The significance of the observed Higgs boson peak is estimated with the procedure presented in Chapter 4. At $m_H = 125.09 \text{ GeV}/c^2$, which is the LHC Run I combined measurement of the Higgs boson mass [13], the observed significance is $10.8\sigma$, for an
In order to extract the signal strength modifier from selected events, a multi-dimensional fit is implemented relying on two variables: the four-lepton invariant mass \( m_{4\ell} \) and the kinematic discriminant \( D_{\text{kin}}^{\text{bkg}} \). A two-dimensional likelihood function is defined as:

\[
L_{2D}(m_{4\ell}, D_{\text{kin}}^{\text{bkg}}) = L(m_{4\ell})L(D_{\text{kin}}^{\text{bkg}}, m_{4\ell})
\] (5.11)

where the \( L(m_{4\ell}) \) likelihood is unbinned and uses the statistical model described in Chapter 4, while \( L(D_{\text{kin}}^{\text{bkg}}, m_{4\ell}) \) is a two-dimensional template of mass vs. the kinematic discriminant. The conditional term of this factor is implemented in the template by normalizing all columns corresponding to the same mass to 1. Therefore, the two-dimensional template does not include any information on the mass, but given a mass value, it provides information on the expected distribution of the kinematic discriminant. The two-dimensional templates are built separately for the three final states (4e, 4µ and 2e2µ) from MC simulations. Based on the seven event categories and the three final states the \((m_{4\ell}, D_{\text{kin}}^{\text{bkg}})\) unbinned distributions are split into 21 channels. A simultaneous fit to all channels is then performed in order to extract the signal strength modifier. The relative fraction of signal events per final state is fixed to the SM prediction and the systematic uncertainties are included as nuisance parameters. The result is obtained using an asymptotic approach with a test statistic based on the profile likelihood ratio Eq. 4.20. The individual contributions of statistical and systematic uncertainties on the final result are separated and to determine the statistical uncertainty a likelihood scan is performed removing the systematic uncertainties. The systematic uncertainty is then taken as the difference in quadrature between the total uncertainty and the statistical uncertainty.

The signal strength result for a fixed mass hypothesis of \( m_H = 125.09 \text{ GeV}/c^2 \) for the inclusive event sample is \( \mu = 1.05^{+0.15}_{-0.14} \text{ (stat)}^{+0.11}_{-0.09} \text{ (syst)} = 1.05^{+0.19}_{-0.17} \). A compatible result is obtained using events belonging to each category separately as shown in Figure 5.10 (top left). The observed values are consistent with the SM prediction of \( \mu = 1 \) within their uncertainties.

In order to extract information on the Higgs boson production mechanisms, five signal strength modifiers are introduced (\( \mu_{ggH}, \mu_{VBF}, \mu_{VH\text{had}}, \mu_{VH\text{lep}} \) and \( \mu_{t\bar{t}H} \)) as scale factors of the expected SM cross section for the different Higgs boson production modes (Section 1.4.1). The WH and ZH processes are merged, and then are splitted in base on the decay of the associated vector boson into either hadronic decays (VHhad) or leptonic decays (VHlep). In order to extract these results a fit is performed similarly to the previous case, but taking into account the expected composition of the different categories in terms of these five Higgs boson production modes which have to be measured. The results obtained are presented in Figure 5.10 (top right). In Table 5.6 the obtained results are compared to the expected signal strength modifiers, where the expected uncertainties are evaluated by generating pseudo-experiments with an Asimov dataset.

78
Figure 5.10: Top left: observed values of the signal strength modifier for the seven event categories, compared to the combined $\mu$ shown as a vertical line. Top right: results of likelihood scans for the signal strength modifiers corresponding to the main SM Higgs boson production modes, compared to the combined $\mu$ shown as a vertical line. The horizontal bars and the filled band indicate the uncertainties within one $\sigma$. The uncertainties include both statistical and systematic sources. Bottom: result of the two-dimensional likelihood scan for the $\mu_{ggH,tH}$ and $\mu_{VBF,VH}$ signal strengths. The solid and dashed contours show the 68% and 95% CL regions, respectively. The cross indicates the best fit values, and the diamond represents the expected values for the SM Higgs boson $[1]$. 

<table>
<thead>
<tr>
<th>Event Categories</th>
<th>$\mu_{ggH,tH}$</th>
<th>$\mu_{VBF}$</th>
<th>$\mu_{VH\text{had}}$</th>
<th>$\mu_{VH\ell p}$</th>
<th>$\mu_{tH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusive</td>
<td>$1.00_{-0.04}^{+0.08}$ (stat) $-0.08$ (syst)</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
</tr>
<tr>
<td>Expected</td>
<td>$1.00_{-0.04}^{+0.08}$ (stat) $-0.08$ (syst)</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
<td>$1.00_{-0.01}^{+0.01}$</td>
</tr>
<tr>
<td>Observed</td>
<td>$1.05_{-0.09}^{+0.09}$ (stat) $0.09$ (syst)</td>
<td>$1.20_{-0.12}^{+0.12}$</td>
<td>$0.05_{-0.05}^{+0.05}$</td>
<td>$0.00_{-0.00}^{+0.00}$</td>
<td>$0.00_{-0.00}^{+0.00}$</td>
</tr>
</tbody>
</table>

Table 5.6: Expected and observed signal strength modifiers $[1]$. 

79
The low observed signal strengths for the VBF, VH, and ttH processes can be explained by the little excess in the untagged category which leads to an observed signal strength higher than the expected value for the ggH process. This higher value contributes significantly to the total signal yield in categories that are based on the hadronic activity. In categories not based on hadronic event activity, events with \( m_4 \ell \sim 125 \text{ GeV}/c^2 \) present a low value of \( D_{\text{bkg}}^{\text{kin}} \) and therefore are more compatible with the background than the signal hypothesis.

In order to study the Higgs boson couplings with fermions and bosons, two additional signal strength modifiers \( \mu_{\text{ggH},t\bar{t}H} \) and \( \mu_{\text{VBF},VH} \) are introduced as scale factors to the expected SM cross section for the fermion and vector-boson induced contribution, respectively.

The measurement of these variables is obtained performing a two-parameter fit simultaneously in all categories, fixing the mass value to the combined Run I result (\( m_H = 125.09 \text{ GeV}/c^2 \)). The results obtained are \( \mu_{\text{ggH},t\bar{t}H} = 1.19^{+0.21}_{-0.20} \) and \( \mu_{\text{VBF},VH} = 0.00^{+0.81}_{-0.00} \) and are shown in Figure 5.10 (bottom) together with the 68% and 95% CL contours. The SM prediction of \( \mu_{\text{ggH},t\bar{t}H} = 1 \) and \( \mu_{\text{VBF},VH} = 1 \) is contained within the 68% CL region.

### 5.6 Study on an alternative event categorization

The event categorization procedure used in the 2016 \( H \rightarrow ZZ \rightarrow 4\ell \) analysis and presented in Section 5.3 is built on a series of subsequent identification criteria based on cut over the physical observables defined in Section 5.2. This categorization has been obtained defining the categorization criteria in empiric way in order to maximize the purity of the tagged Higgs boson production mode, without penalize the number of event in each category.

A study on an alternative event categorization has been performed and the results obtained are reported in this section.

The purpose of this study is to improve the sensitivity of the analysis to the Higgs boson production modes by maximizing the purity of each category, i.e. the number of identified events in each category. This is obtained with a machine learning algorithm which uses the same event information as the current categorization in order to classify events, i.e. the number of extra leptons, the number of jets and b-jets and the MELA discriminants \( D_{2\text{jet}}, D_{1\text{jet}}, D_{WH}, D_{ZH} \).

The difference with the current categorization method is that the categorization criteria are not based on fixed cut of the input variables, but on an automatic classificator trained on Monte Carlo events. With this technique an ensemble of Machine Learning models are trained in order to classify the events of a certain dataset in categories with successive random extraction and in addition after each training iteration the events get reweighed in order to recollect events that were incorrectly classified by the previous models. In the end the predictions of all models in the ensemble are combined in order
Chapter 5. Measurement of Higgs boson properties with 2016 data

Figure 5.11: Relative signal purity in the seven event categories of the alternative categorization study in term of the six considered the Higgs boson production mechanisms described in the text [41].

to get a common prediction. This algorithm has been implemented using two boosting algorithms, Adaboost and Extreme Gradient Descent Boosting. Further details on the Machine Learning technique used can be found in Ref. [41].

This alternative categorization was optimized for seven categories: ggH-tagged, VBFH-tagged, VH-lept-tagged, VH-hadr-tagged, ttH-tagged, bbH-tagged and ZH-met-tagged. The main differences with respect to the standard categorization (defined in Section 5.3) are:

- A single category, VBFH-tagged, targets VBF events, instead of two categories (VBF-1jet-tagged and VBF-2jet-tagged) which divided VBF events according to the number of additional jets;
- The bbH-tagged category is added in order to take into account the associated production of the Higgs boson with a pair of bb quarks;
- The ggH events are collected in a dedicated ggH-tagged category, instead of being left in the untagged category;
- A ZH-met category is introduced, specifically targeting events with associated production of a Z boson decaying to neutrinos.
Figure 5.12: Expected valued of the signal strength modifiers corresponding to the Higgs boson production modes obtained with the alternative categorization (left) and with the categorization currently used in the $H \rightarrow ZZ \rightarrow 4\ell$ analysis (right), removing the systematic uncertainties in order to allow a comparison between the two results. The central values are fixed to the SM expectation and the horizontal bars indicate the uncertainties within one $\sigma$.

The performance of this alternative categorization is presented in Figure 5.11, which shows that this categorization method can produce categories with high purity. In order to verify if the enhancement of signal purity in categories improve the analysis sensitivity, it is necessary to test if the uncertainties on expected yields are reduced.

In order to extract the expected yields and expected signal strength modifiers for this alternative categorization a statistical analysis is performed using the statistical model defined in Chapter 4.

The inputs of the statistical model used for the analysis are the similar to the ones presented in the Section 4.5 but with some differences:

- Expected signal yields are computed from MC samples with the machine learning technique used to categorize events only for the mass point $m_H = 125 \text{ GeV}/c^2$;
- Only the $q\bar{q} \rightarrow ZZ$ process is considered as background, and its expected yield is extracted form MC;
- Expected signal shapes are described by a double crystal ball function (Eq. 5.9) with the same parameters in all the production modes and categories considered;
- Category migration systematic uncertainties are not considered, like for example systematic uncertainties on the jet energy scale or on the b-tagging efficiency whose effect is to change the number of jets or b-jets in one event causing its migration from one category to another.
The analysis has been performed generating pseudo-experiments with this statistical model using the alternative categorization and extracting the expected signal strength modifiers per production mode, shown in Figure 5.12 (left). These results are compared with the expected signal strength modifiers extracted with this same statistical model from the current $H \rightarrow ZZ \rightarrow 4\ell$ analysis event categorization, shown in Figure 5.12 (right), obtained using the same simplified model as the alternative categorization in order to allow the comparison of the results.

The result shows that despite the better purity of alternative categories, there is no improvement in the uncertainties on the signal strengths per production mode. This can be explained by the fact that the larger purity is obtained at a cost of significant lower yield, which entails larger statistical uncertainties.

This study on an alternative event categorization showed that in order to optimize the performance of event categorization, maximizing the purity of each category is not sufficient, on the contrary it is necessary to use other criteria which account also the statistical significance of each Higgs boson production mechanism in the corresponding category.

The result obtained from this study will be used in future studies for the optimization of the event categorization.
Chapter 6

A first look at 2017 data

This chapter presents the first quality checks performed on the first 2017 data. In May 2017 the LHC started again delivering proton-proton collisions, after the winter-stop in which some sub detectors (like for example the pixel detector) have been upgraded. The data taking now is ongoing and the sample with which the first checks are performed corresponds to an integrated luminosity of 13.88 fb$^{-1}$.

The controls performed on the first 2017 sample are fundamental in order to ensure the quality of collected data and to check possible issues with the detectors, especially since the CMS pixel detector has been upgraded, and since trigger menus have been updated to cope with the increase of instantaneous luminosity. These checks allow early feedback to be provided to the sub detector systems.

First data checks are performed considering the distributions of the reconstructed quantities relevant to the analysis (lepton scale, lepton selection variables or Z boson reconstruction rate) and their stability in time. The obtained results are then compared with 2016 MC simulated samples and with 2016 data.

The purpose of these controls is to prepare the environment for the analysis which will be performed with the full 2017 data sample, once the data taking is complete at the end of the year.

6.1 Data and MC samples

6.1.1 2017 data sample

Controls are performed on the proton-proton collision data sample recorded by the CMS detector in the first data taking months (from May to September) of 2017, corresponding to an integrated luminosity of 13.88 fb$^{-1}$. Collision events are selected by the same high-level trigger algorithms used for 2016 data analysis (see Table 3.1) with the addition of the HLT paths presented in Table 6.1. These new triggers have been added since with the increase of the on-line luminosity some HLT path, especially singleElectron paths, have to be rescaled.

As for 2016 data, the trigger algorithm selects events which present leptons passing loose
Chapter 6. A first look at 2017 data

<table>
<thead>
<tr>
<th>HTL path</th>
<th>PDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiEle</td>
<td>DoubleEG</td>
</tr>
<tr>
<td>SingleEle</td>
<td>SingleElectron</td>
</tr>
<tr>
<td>HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele35_WPTight_Gsf</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele38_WPTight_Gsf</td>
<td></td>
</tr>
<tr>
<td>HLT_Ele40_WPTight_Gsf</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Trigger paths added to those showed in Table 3.1 to be used in 2017 collision data, and their corresponding primary datasets.

identification and isolation requirements, described in Chapter 3. The events which pass the trigger selection are required to pass the \( H \rightarrow ZZ \) event selection criteria used for the 2016 analysis and described in Section 3.8.2.

6.1.2 2016 data and MC samples

The results obtained with the 2017 data sample are compared with the ones extracted from 2016 data and MC.

The 2017 MC samples are not yet available since they have to include the simulation of the upgraded pixel detector and all details of the 2017 data taking conditions. Therefore the results extracted from 2017 data are compared with 2016 MC samples described in Section 5.1. In particular for the first check results the comparison is done using the Drell-Yan and \( t \bar{t} \) samples.

The 2016 data sample considered is the one corresponding to the full 2016 dataset used to perform the analysis illustrated in Chapter 5. It corresponds to proton-proton collision data collected in 2016 by the CMS detector up to an integrated luminosity of 35.9 fb\(^{-1}\). Events selected by the 2016 triggers (Table 3.1) are required to pass the \( H \rightarrow ZZ \) event selection criteria used for the 2016 analysis and described in Section 3.8.2.

6.2 Event selection

After trigger selection, events are required to pass the event selection criteria used for the 2016 analysis to be considered for performing the first quality checks on 2017 data. The events considered are not only four-lepton signal events, but according to the kind of analysis check performed, different types of events are used. The classes of events used are:

- **Z events**, which include a Z candidate for each reconstructed event. The Z candidate is built from a pair of selected leptons with 2016 \( H \rightarrow ZZ \) analysis criteria described in Section 3.8.2. If one event contains more than one Z candidate, the one with the closer reconstructed invariant mass to the nominal Z boson mass is selected.

- **Z+L events**, which include a Z candidate plus exactly one lepton. The Z candidate is built from a pair of leptons selected with 2016 analysis criteria described
in Section 3.8.2 while the additional lepton is required to pass the loose selection requirements for leptons. These events form the $H \rightarrow ZZ \rightarrow 4\ell Z\text{+L}$ control region that is used to estimate the lepton fake rate.

- **ZZ events**, which include a ZZ candidate selected with the procedure described in Section 3.8.2

## 6.3 Corrections applied

Since the data taking started in May, updated lepton energy scale corrections as those described in Section 3.5 are not yet available and will be derived at the end of the data taking. Therefore for the 2017 data sample the electron momentum estimated in the “prompt”reconstruction is used, with no additional correction. For the muons instead the 2016 scale corrections were found to marginally improve the reconstructed Z lineshape and the data-MC agreement and are therefore included.

The 2016 data sample considered contains all the corrections provided for the analysis illustrated in Chapter 5.

Since 2017 MC are not yet available, 2016 MC samples are used. But when 2017 data are compared with 2016 MC it is necessary to reweigh MC events with an appropriate weight which accounts for the event pile-up (number of overlapping proton-proton interactions) in 2017 data, in order to match the number of interactions per LHC bunch crossing observed in 2017 data with the MC simulations used.

Figure 6.1 shows the distribution of the number of proton-proton interactions per LHC bunch crossing. The pile-up weight is obtained from the ratio between data and MC distributions. Even if the average of the 2017 data distribution moves to higher values, it is still possible to model the 2017 the pile-up correction with 2016 MC thanks to the tail present in the MC distribution.

## 6.4 Data - MC comparison

The fist check performed is the comparison between 2017 data and MC, in order to test the agreement between the two samples and to verify the necessity of scale and efficiency corrections.

The analysis is performed considering the Z boson di-lepton invariant mass peak and dividing the Z decays in $e^+e^-$ and $\mu^+\mu^-$ in order to separate the contribution from electrons and muons whose momentum scale and inefficiencies derive from different sources. The 2017 data sample corresponding to $13.88 \text{ fb}^{-1}$ is compared to the 2016 MC Drell-Yan and $t\bar{t}$ samples, reweighed (see Section 6.3) in order to match the data distribution.

The analysis is firstly performed in the Z event sample and the results obtained are shown in Figure 6.2 for the Z boson decaying in electrons (left frame) and in muons.
Chapter 6. A first look at 2017 data

Figure 6.1: Distribution of number of proton-proton interactions per LHC bunch crossing for the 2016 MC sample (solid histogram), 2016 data (black points) and 2017 data (red points).

(right frame). The bottom pad of the two plots shows the ratio between the data and MC histograms.

The agreement between data and MC is found to be quite good for muons. For electrons, the scale discrepancy between data and MC is found to be of 0.3% which is of the same order of the systematic uncertainty estimated on the electron energy scale (see Section 5.4.3). Discrepancies like this are expected for electrons since they rely on the ECAL measurements for the determination of their energy, and can be corrected \textit{a posteriori} introducing the scale corrections described in Section 3.5, which will be determined at the end of the data taking.

Data and MC distributions are then fitted with a Double Crystal Ball (DCB) function in order to quantify the data-MC agreement. The parameters of the DCB function obtained from the fit are reported in Table 6.2 and show the good agreement between data and MC.

This study has been performed also considering leptons reconstructed in different regions of the CMS detector, in order to check if the discrepancy between data and MC can be due to difference in event reconstruction or selection in the different detector regions. Figure 6.3 shows the data-MC comparison for events of the Z sample, where the Z boson decays into $e^+e^-$ which are reconstructed in the ECAL barrel (left plot) or in the ECAL endcaps (right plot). From the plots it is possible to notice that the already mentioned discrepancy is mostly for endcap electrons. This is due to larger amount of radiation collected by the ECAL endcaps with respect to the barrel. This causes more stringent
Chapter 6. A first look at 2017 data

Figure 6.2: Comparison between 2017 data corresponding to an integrated luminosity of 13.88 fb$^{-1}$ and 2016 MC invariant mass distributions of the $Z$ boson, decaying in $e^+e^-$ (left frame) and $\mu^+\mu^-$ (right frame), in the $Z$ events sample. The bottom pads represent the ratio between data and MC histograms.

Table 6.2: Mean and width parameters of the Double Crystal Ball functions used to fit the Data and MC di-lepton invariant mass distributions of the $Z$ events sample, separately for $Z \rightarrow e^+e^-$ and $Z \rightarrow \mu^+\mu^-$.  

<table>
<thead>
<tr>
<th></th>
<th>$Z \rightarrow e^+e^-$</th>
<th>$Z \rightarrow \mu^+\mu^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>90.15</td>
<td>2.97</td>
</tr>
<tr>
<td>MC</td>
<td>90.51</td>
<td>2.78</td>
</tr>
</tbody>
</table>

requirements for the detector calibration and in lepton scale corrections. In particular the reconstructed $Z$ peak is shifted in data with respect to MC, and therefore it will be necessary to provide updated scale corrections for electrons. Table 6.3 reports the DCB parameters obtained from the fit of the data and MC distributions in order to quantify the data-MC agreement.

The same distributions obtained for $Z \rightarrow \mu^+\mu^-$ events present a good data-MC agreement also for muons reconstructed in the CMS muon detector endcaps, and this is also thanks to the muon scale corrections already included.

This analysis has also been performed on the $Z+L$ event sample, in order to control the data-MC agreement also in a region more similar to the $ZZ$ signal region, which is the interest of the $H \rightarrow ZZ \rightarrow 4\ell$ analysis. The comparison between data and MC is performed considering the $Z$ boson reconstructed in $Z+L$ events. The results obtained
Figure 6.3: Comparison between 2017 data corresponding to an integrated luminosity of 13.88 fb$^{-1}$ and 2016 MC invariant mass distributions of the Z boson decaying in $e^+e^-$ for the Z events sample, with electrons both reconstructed in the ECAL barrel (left plot) and in the ECAL endcaps (right plot). The bottom pads represent the ratio between data and MC histograms.

for this sample are similar to the one already presented for the Z event sample. As for Z events the agreement between data and MC is found to be better for muons than for electrons (Figure 6.4, Table 6.4). The distributions obtained for Z+L events present larger error bars than the ones in the Z event sample and this is due to the lower number of events selected in the Z+L event sample. However, the observed results for the Z+L sample present the same level of agreement between data and MC as for the Z event sample.

In conclusion 2017 lepton scale and resolution are found to be in reasonable agreement with 2016 MC despite some discrepancies for electrons which will be accounted by energy scale corrections, to be derived on the full 2017 data set at the end of the run.

6.5 Time dependency study

A second check performed on the 2017 data is the study of the evolution in time of some quantities relevant to the analysis. The analysis is done on the Z and Z+L classes of events (Section 6.2) and it is performed considering the reconstructed Z candidate in these samples. The quantities of interest can be collected in three groups: Z boson properties (i.e. Z mass, Z width), Z boson reconstruction rate and lepton properties (i.e. lepton isolation or SIP), considering the leptons coming from the Z boson decay.
Table 6.3: Mean and width parameters of the Double Crystal Ball functions used to fit the Data and MC di-lepton invariant mass distributions of the Z events ample, with the Z boson decaying in $e^+e^-$, separately for electron reconstructed in the ECAL barrel and in the ECAL endcaps.

<table>
<thead>
<tr>
<th></th>
<th>$e^+e^-$ in Barrel</th>
<th>$e^+e^-$ in Endcap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>90.70 [GeV/c$^2$]</td>
<td>89.14 [GeV/c$^2$]</td>
</tr>
<tr>
<td>MC</td>
<td>90.52 [GeV/c$^2$]</td>
<td>90.295 [GeV/c$^2$]</td>
</tr>
</tbody>
</table>

Figure 6.4: Comparison between 2017 data corresponding to an integrated luminosity of 13.88 fb$^{-1}$ and 2016 MC invariant mass distributions of the Z boson, decaying in $e^+e^-$ (left frame) and $\mu^+\mu^-$ (right frame), in the Z+L events sample. The bottom pads represent the ratio between data and MC histograms.

This time dependency analysis is performed in order to verify the stability in time of the quantities of interest. Variations in time of these quantities derive from any corrections or variations during the data taking. When data are collected by CERN detectors, they are stored together with the characteristics of the trigger requirements and the detector set-up in which the data taking is performed. The change in some parameters (such as an update in a trigger menu) determines the end of the data taking period and the beginning of a new one. Each period is identified by a letter and can be divided in sub-periods due to minor changes (sub-periods are identifies by a number). The 2017 data sample considered, which corresponds to an integrated luminosity of 13.88 fb$^{-1}$, is divided in two data taking periods (B and C) and their sub-periods. Table 6.5 reports the name of the data taking period and the corresponding integrated luminosity collected in each of them.
Chapter 6. A first look at 2017 data

<table>
<thead>
<tr>
<th></th>
<th>$Z \rightarrow e^+e^-$</th>
<th></th>
<th>$Z \rightarrow \mu^+\mu^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>90.32</td>
<td>2.93</td>
<td>90.96</td>
</tr>
<tr>
<td>MC</td>
<td>90.65</td>
<td>2.71</td>
<td>90.97</td>
</tr>
</tbody>
</table>

Table 6.4: Mean and width parameters of the Double Crystal Ball functions used to fit the Data and MC di-lepton invariant mass distributions of the Z+L event sample, separately for $Z \rightarrow e^+e^-$ and $Z \rightarrow \mu^+\mu^-$. 

<table>
<thead>
<tr>
<th>Data taking era</th>
<th>L [fb$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017B-v1</td>
<td>3.873</td>
</tr>
<tr>
<td>2017B-v2</td>
<td>0.680</td>
</tr>
<tr>
<td>2017C-v1</td>
<td>1.277</td>
</tr>
<tr>
<td>2017C-v2</td>
<td>4.043</td>
</tr>
<tr>
<td>2017C-v3</td>
<td>4.007</td>
</tr>
</tbody>
</table>

Table 6.5: Data taking periods (eras) and corresponding integrated luminosity (L) collected in each of them.

Since as time passes more and more data are collected, the amount of recorded integrated luminosity can express the time evolution. Therefore in the performed analysis the variables are studied looking at their distribution with respect to the integrated luminosity recorded.

The analysis is performed by dividing collected data into luminosity blocks of about 0.5 fb$^{-1}$. In each block the distributions of the quantities relevant to the analysis are considered and their parameters are evaluated. The values obtained per luminosity block are then plotted vs. the integrated luminosity.

The evaluation method of distribution parameters is different for the various quantities of interest.

6.5.1 Z boson properties

In order to extract Z boson properties the Z mass distribution is considered in each luminosity block and a fit is performed with a Double Crystal Ball (DCB) function, separately for $Z \rightarrow e^+e^-$ and $Z \rightarrow \mu^+\mu^-$. The resulting DCB mean and width of the Z invariant mass distributions are then plotted vs. the recorded integrated luminosity.

Checking the stability of Z boson reconstructed mass can give information on lepton energy scale, while checks on the width of the Z mass distribution gives information on the lepton resolution.
Chapter 6. A first look at 2017 data

Figure 6.5: Distribution of the DCB mean (left) and width (right) vs. integrated luminosity for the Z event data sample. Points represent the results obtained for muons (red) and for electrons (blue) coming from the Z boson decays. The dashed lines represent the mean of the full 2017 data sample, while the solid lines are the mean of 2016 MC sample.

This analysis is first performed on the Z event sample and the results obtained are reported in Figure 6.5 for the mean (left) and for the width (right) of the Z mass distributions. The reconstructed Z boson mass and width are found to be quite stable with time, even if there are some larger fluctuations for electron decays (\(\sim 1\%\)) both between different eras and within data taking periods. These fluctuations will be corrected with time-dependent scale corrections which will be provided at the end of the run. The distributions vs. luminosity are then compared with the mean value of 2016 MC (solid line) and the agreement data-MC is found to be better for muons than for electrons. The data-MC agreement is key to the reliability of the statistical model and it is actually more important than the absolute correctness of the scale. As it is possible to see from the Z mass plot (Figure 6.5 left) muons almost perfectly agree with the MC prediction while the electrons agrees with MC within 1% variation. Also for the Z boson width (Figure 6.5 right) it is possible to notice a better data-MC agreement for muons. In particular in 2017 data the Z width is larger than the 2016 result due to a worse lepton resolution. This is expected to be improved with lepton scale corrections. The analysis is also performed considering the different detector regions where leptons coming from the \(Z \rightarrow e^+e^-\) and \(Z \rightarrow \mu^+\mu^-\) decays are reconstructed. The distribution obtained are found to be quite stable in time, with the major discrepancies coming from leptons reconstructed in the endcap regions.

This analysis then is performed on the Z+L event class in order to probe a data region more similar to the ZZ signal region. The obtained results are shown in Figure 6.6 for the mean (left) and for the width (right) of the Z mass distributions. The obtained trend...
Figure 6.6: Distribution of the DCB mean (left) and width (right) vs. integrated luminosity for the Z+L event data sample. Points represent the results obtained for muons (red) and for electrons (blue) coming from the Z boson decays. The dashed lines represent the mean of the full 2017 data sample, while the solid lines are the mean of 2016 MC sample.

is quite similar to the one obtained analysing Z events, only with larger error bars due to the lower statistics in the Z+L sample with respect to Z sample. Also in this case the distributions vs. luminosity are compared with the mean value of 2016 MC (solid line) and the results observed are similar to those for the Z event sample.

6.5.2 Z boson reconstruction rate

Another check performed on the 2017 data sample is checking the rate of the reconstructed Z bosons. This quantity can give information on the trigger efficiency and on the lepton selection and reconstruction efficiency.

The analysis is performed dividing in luminosity blocks the 2017 data sample, first considering the Z event sample and then on the Z+L sample. In order to estimate the Z production rate, in each luminosity block the number of events in the Z boson mass distribution is divided for the integrated luminosity of the corresponding block, separately for $Z \rightarrow e^+e^-$ and $Z \rightarrow \mu^+\mu^-$. The values obtained are then plotted vs. the integrated luminosity and the results are shown in Figure 6.7 for Z events (left plot) and for Z+L events (right plot).

The Z rate estimated for the 2017 data sample is then compared with the mean value obtained from 2016 data collected in the last 2016 data taking period (solid lines in the plots) in order to test possible changes on lepton selection and reconstruction efficiency. It is possible to notice that the Z production rate is reasonably stable with time both in Z and in Z+L events and also their trends are similar. The rate is similar to that of 2017 for the Z sample, which is an indication of similar efficiency for prompt leptons. However, a lower rate of Z bosons decaying into muons (red points) is observed in the Z+L sample (right plot). This difference from 2016 data can be partly explained by
Chapter 6. A first look at 2017 data

Figure 6.7: Distribution of the Z boson reconstruction rate for the Z (left plot) and Z+L (right plot) event data samples. Points represent the results obtained for muons (red) and for electrons (blue) coming from the Z boson decays. The solids lines represent the Z boson production rate mean obtained from 2016 data collected in the last 2016 data taking period.

variations in lepton isolation and SIP distributions which can be due to the upgrade of the pixel detector of the CMS experiment. The decrease of the Z+L rate while the Z rate is constant can be attributed by a variation of efficiency for the additional lepton, which is often non-prompt, seen the definition of the control region Z+L. Therefore this can be an indication of a smaller number of potential fake leptons, but further studies are necessary in order to verify this hypothesis. This result suggests that a study on selection variables is necessary in order to test the level of comprehension of this phenomenon, in order to accurately describe it with the MC samples and if necessary re-tune lepton selection cuts. The study of the lepton selection variables is presented in the following section.

6.5.3 Lepton properties

In order to study the possible variations of isolation and SIP distributions in 2017 data and to check for potential issues with these variables that are used to select leptons in the analysis the time-dependency study is performed for the Z+L sample. This study is performed only on this sample since it presents a phase space more similar to the one of the ZZ signal region and an additional loose lepton. In order to extract the isolation results in the Z+L sample, the leptons coming from the Z boson decays which present the maximum (worst) value of isolation is considered. In each luminosity block, in which the data sample is divided, the distribution of these maximum values of isolation is built, and then the mean values of all the distributions are plotted vs. the integrated luminosity. The same analysis is repeated for the maximum
lepton SIP, always considering leptons coming from Z boson decays.

The obtained result is shown in Figure 6.8 for maximum isolation (left) and maximum SIP (right), separately for $\text{Z} \rightarrow e^+e^-$ and $\text{Z} \rightarrow \mu^+\mu^-$. The 2017 results are compared with the same quantities obtained with 2016 data from the last 2016 data taking period, represented in the plots as solid lines. It is possible to notice that the isolation distribution is quite flat and in agreement with the 2016 mean value, while the SIP distribution presents some discrepancies. In particular the maximum SIP distribution for electrons in 2017 data is significantly lower with respect to the 2016 mean value. This fact may be correlated with the additional layer of the pixel detector added in the CMS tracking system. This additional layer, being the nearest to the beam pipe, improves the determination of the track impact parameter. This observed difference will be further investigated also to test if the 2017 MC simulation, which is currently being produced, will correctly reproduce the SIP distributions. However, since the cut currently used in the analysis is quite loose ($\text{SIP} < 4$), a small variation on the signal prompt-lepton selection efficiency is expected. Nevertheless, the amount of expected background could change and a study will be necessary in order to test the necessity of re-tune the SIP cut.

### 6.6 Effect of FSR recovery

Since leptons from Z boson decays can present final state radiation (FSR), an FSR recovery algorithm is used to recollect FSR photons in order to better reconstruct the invariant mass of the initial particle (Section 3.4). Since potential issues with photon reconstruction and identification can affect the result, a further study performed on first 2017 data is focused on FSR events, in order to
Figure 6.9: Z boson invariant mass distributions without (red line) and with (blue line) including FSR photons in the mass computation, for 2017 data (left) and 2016 data (right).

<table>
<thead>
<tr>
<th>Decay</th>
<th>2016 data</th>
<th>2017 data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z \rightarrow \ell^+\ell^-$</td>
<td>3.36%</td>
<td>2.92%</td>
</tr>
<tr>
<td>$Z \rightarrow e^+e^-$</td>
<td>2.51%</td>
<td>2.17%</td>
</tr>
<tr>
<td>$Z \rightarrow \mu^+\mu^-$</td>
<td>3.96%</td>
<td>3.53%</td>
</tr>
</tbody>
</table>

Table 6.6: Fractions of Z events with FSR in 2016 and 2017 data samples, for the inclusive Z boson decay and separately for $Z \rightarrow e^+e^-$ and $Z \rightarrow \mu^+\mu^-$. 

check the number of events with FSR and how the invariant mass reconstruction differs including or not FSR photons. This study can give information on the efficiency of the FSR recovery algorithm.

The study is performed on the Z event sample reconstructing the Z boson invariant mass of events which presents FSR photons, first not considering the FSR photon in the invariant mass reconstruction and then including FSR. The results obtained are shown in Figure 6.9 comparing 2017 data outcome (left plot) with 2016 results (right plot). The reconstructed invariant mass distributions are similar in 2017 and 2016 data, and also the fractions of FSR events over the total number of Z candidates (shown in Table 6.6) are compatible.
Table 6.7: Number of expected and observed background and signal events for each final state, for the $122 < m_{4\ell} < 128 \text{ GeV/c}^2$ mass range, for an integrated luminosity of 13.88 $\text{fb}^{-1}$. Irreducible backgrounds and signal yields are estimated using 2016 Monte Carlo samples. The reducible background prediction is obtained scaling the prediction for 2016 data according to the integrated luminosity. The uncertainties reported for the reducible backgrounds and signal include only the statistical uncertainties on the MC samples dimensions, while the systematic uncertainties on these predictions are those discussed in Section 5.4.3. The uncertainties reported for the $Z+X$ prediction include only systematic uncertainties on the reducible background estimation (Section 5.4.3).

### 6.7 ZZ mass distributions

The first checks performed on events not contained in the ZZ signal region showed that there are not fundamental problems in event selection or reconstruction efficiency, although some corrections are needed in order to achieve a better data-MC agreement. Such corrections are routinely produced at the end of every data taking period, once they are validated over the full data sets, and will therefore be adopted for the full Run II analysis. In the meanwhile, it is possible to look at the ZZ signal region. This section presents the results obtained with the 2017 data ZZ event sample. The ZZ events are selected with 2016 $H \rightarrow ZZ \rightarrow 4\ell$ analysis criteria presented in Section 3.8.2.

Figure 6.10 shows the four-lepton invariant mass distribution obtained with first 2017 data sample corresponding to an integrated luminosity of 13.88 $\text{fb}^{-1}$ both for the full mass range (top plot) and in the region around the Higgs boson peak (bottom plot). The Higgs boson appears as expected as a narrow peak on a flat background distribution. The observed data are compared with the 2016 MC samples. The normalization of the SM Higgs boson signal and the ZZ backgrounds are estimated according to the respective cross-sections. Since the normalization of the $Z+X$ background must be estimated from data and the estimation corresponding to 2017 data is not yet available, the reducible background normalization is obtained rescaling the 2016 prediction according to the luminosity.

It is possible to notice the good agreement between data and MC samples around the
Higgs boson peak and in the $Z \rightarrow 4\ell$ boson peak region. In the region on the right of the Higgs boson peak, however, an excess can be noticed in observed data with respect to the expected background. This is due to the fact that the reducible background is estimated only by rescaling the 2016 prediction according to the 2017 data sample luminosity. Seen the variations of the lepton variables (i.e. SIP) and of the $Z+L$ rate, as discussed in the previous sections [6.5.2] and [6.5.3] a complete estimation of the reducible background will be necessary on the basis of 2017 data.

Figure [6.11] shows the four-lepton invariant mass distribution separately for the three final states $4e$, $4\mu$ and $2e2\mu$. Also in these plots it is possible to notice the good agreement between data and MC in the Higgs boson peak and $Z$ boson peak, especially for the $4\mu$ final state.

Table [6.7] shows the expected and observed signal and background yields obtained in the $122 < m_{4\ell} < 128$ GeV/$c^2$ mass range. The observed yields are compatible with the expectation, with some excess mainly driven by the $4e$ channel. The agreement is however satisfactory considering the current knowledge of reducible background, as described before.

A complete statistical analysis as the one described in Chapter [5] will be possible once appropriate MC simulations of the 2017 data taking conditions are available and will require all lepton scale corrections and the full background estimation. However, this initial study shows a good agreement between the observed Higgs boson peak and the MC expectation.

The agreement between data and MC is shown also by the reconstructed mass of the $Z_1$ and $Z_2$ bosons, shown in Figure [6.12] and by the kinematic discriminant $D_{bkg}^{\text{kin}}$ (discussed in Section [5.2.1]), shown in Figure [6.13].
Figure 6.10: Distribution of the reconstructed four-lepton invariant mass $m_{4\ell}$ in the full mass range (top frame) and the low-mass range (bottom frame) obtained with an early 2017 data sample corresponding to an integrated luminosity of 13.88 fb$^{-1}$. 
Figure 6.11: Distribution of the reconstructed four-lepton invariant mass $m_{4\ell}$ in the full mass range in the three final states $4e$ (top left), $4\mu$ (top right) and $2e2\mu$ (bottom) obtained with an early 2017 data sample corresponding to an integrated luminosity of $13.88 \text{ fb}^{-1}$.
Figure 6.12: Distribution of the $Z_1$ (left) and $Z_2$ (right) reconstructed invariant masses in the full mass region obtained with an early 2017 data sample corresponding to an integrated luminosity of 13.88 fb$^{-1}$. The stacked histograms represent the expected signal and background distributions, and points represent the data.

Figure 6.13: Distribution of the kinematic discriminant $D_{\text{kin}}^{\text{bkg}}$ obtained with an early 2017 data sample corresponding to an integrated luminosity of 13.88 fb$^{-1}$. The stacked histograms represent the expected signal and background distributions, and points represent the data.
Conclusions

After the Higgs boson discovery in 2012, further studies performed on this particle are aimed at investigating its nature and test possible deviations from the Standard Model predictions. The study of Higgs boson sub-dominant production mechanisms is carried in this context. These production modes present a cross section more than one order of magnitude lower with respect to the dominant Higgs boson production mode, the gluon fusion mechanism, and therefore are difficult to be observed. With the increase of the integrated luminosity collected by LHC detectors, these processes become accessible.

This thesis focuses on the analysis of the $H \rightarrow ZZ \rightarrow 4\ell$ process in order to measure Higgs boson properties. The analysis presented is performed with LHC Run II data and it is part of the broader analysis effort of the CMS Collaboration on this “golden channel”.

The first part of this thesis presents the analysis performed on 2016 data collected by the CMS Experiment, firstly introducing the event selection and reconstruction procedure, then illustrating the statistical model needed in order to extract results, and finally presenting these results, which were published in a paper [1]. I contributed to this analysis by preparing the statistical model (Chapter 4), collecting and validating all inputs, i.e. expected yields, invariant mass lineshapes, and systematic uncertainties, like for example the systematic uncertainty due to the modelling of lepton efficiency in MC (Section 5.4). Then I performed the statistical analysis with the defined statistical model and I produced the results in term of signal strength modifiers. These quantities are introduced as scale factors of the Standard Model prediction for the Higgs boson cross section, inclusive and per production mode. The results obtained are found to be compatible within the uncertainties with the SM predictions (Section 5.5).

Although the observed results are compatible with the predictions, they can improve with higher luminosity collected. The analysis is set up in this perspective, dividing collected events in categories (Section 5.3) sensitive to the Higgs boson production mechanisms, in order to measure their cross sections. The event categorization is performed according to successive cuts on physical observables (Section 5.2) which characterize particular types of events. It is optimized in order to maximize the signal purity without suppressing the statistics in each category. The current event categorization is in fact ambitious for the available integrated luminosity: some categories present very low number of expected events. The improvement of the statistical technique is therefore a key point for the
future $H \to ZZ \to 4\ell$ analysis which will be performed with the expected luminosity for the end of the LHC Run II ($\sim 150\text{fb}^{-1}$). The aim is to improve the sensitivity to Higgs boson secondary production modes.

In order to study a possible improvement of the event categorization, I performed a study on an alternative categorization method (Section 5.6) based on a machine learning algorithm, which relies on an automatic classifier trained on Monte Carlo events. The obtained categorization, however, maximizes the signal purity in spite of the number of events in each category, which brings larger statistical uncertainties on the results. The outcome of this study is the confirmation that the categorization used in the current analysis is the definition which provides the best sensitivity to the Higgs boson rare production modes.

In May 2017 the LHC started again delivering proton-proton collisions. The extension of the analysis of Higgs boson properties to the full integrated luminosity available is important to reach the maximum possible sensitivity on the Higgs boson production modes. Therefore I performed a preliminary analysis on a first sample of 2017 data, corresponding to an integrated luminosity of $13.88\text{fb}^{-1}$ in the $H \to ZZ \to 4\ell$ channel (Chapter 6). Even if the 2017 data taking is still ongoing, these early studies are necessary to check the quality of collected data and to prepare the environment for the analysis which will be performed with the full 2017 data sample. For this reason I tested the agreement of 2017 data with Monte Carlo simulations and 2016 data, and I performed a time dependency study considering the distributions of some quantities relevant for the analysis in time. These controls aimed at testing the quality of collected data and the performance of detectors in order to verify the necessity of corrections for data and MC. These controls evidenced a quite positive state of the analysis with the good data-MC agreement, despite of the lack of scale factors (e.g. lepton momentum scale calibration) and of updated MC to reflect the 2017 data taking conditions. These will be provided further on for the 2017 analysis.

After these checks, I performed a first analysis on 2017 data observing the Higgs boson peak as a narrow resonance on the four-lepton invariant mass distribution. The observed invariant mass distribution is found to be in agreement with the MC prediction in the Higgs boson peak and $Z$ boson peak regions. The agreement with MC is less satisfactory in the region above the Higgs boson peak. This can be explained by a variation of the contribution of reducible background due to different detector conditions compared to 2016. This will be addressed by a re-optimization of the selection cuts and a proper estimation of the contribution of reducible background once 2017 data taking will be completed. Overall, this early look at the 2017 data shows that there is no major concern, and that the machine and detector are delivering good quality data that will eventually allow a deeper investigation of Higgs boson production mechanisms and properties.
Acknowledgements

I would like to thank Professor Nicola Amapane, my supervisor, for the time and the patience he spent teaching me far more than the technical and scientific contents necessary for the thesis, and for the enthusiasm and the curiosity that made me look at things in a new way.
I would also like to thank my family and friends for the encouragement they have given me during this five-year adventure.
Bibliography


[22] W. Adam, R. Frühwirth, A. Strandlie, and T. Todor, “Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC,”.


