Detector optimisation for future linear collider

Boruo Xu
of King's College

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

Abstract

This is my abstract. To be or not to be.

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university. This dissertation does not exceed the word limit for the respective Degree Committee.

Boruo Xu

Acknowledgements

Of the many people who deserve thanks, some are particularly prominent, such as my supervisor. . .

Preface

This will be my preface. Where is Wolly?

x

Contents

"Two bags of pork scratchings are worth a bag of gold."
  — Joris the Dutch

Chapter 1

Let's make introduction great again

"Introduction means introdcution"
— Theresa Trump

Introduction

Chapter 2

Theory

"ILC will be built next year"
— Mysterious person

overall

ILC

CLIC

calorimeter

ECal

HCal

Muon chamber

Forward detector

Tracker

CLCI-background

Theory

2.1 Motivation

Since Higgs discovery in the LHC in 2012, Higgs

Ha there is a higgs.

We found higgs. Higgs is cool. It explains mass.

Why double higgs. Double higgs coupling is unique to linear collider. It can revel much about the BSM models.

Generator level study has performed. ILC has done this this and that. gHHH in CLIC before

Here we do things differently. First subchannels, then extract both couplings simultaneously.

2.2 Theory

general higgs field

Lagrangian

current constraunt

single higgs coupling measurement done in higgs

Double higgs measurement

The main mechanism for double Higgs production

Chapter 3

Reconstruction

"How to open a pandora box?"
— A wise Chinese

As previously stated, this document focus on the ILD and CLIC_ILD detectors. Due to the similarity, we often only discuss one detector to avoid the repetition. The difference in detectors will be stated if applicable.

3.1 Reconstruction overall

Reconstruction software runs in Marlin software framework [1], part of iLCSoft.

The reconstruction of a event into physical objects containing a few main steps: digitisation of calorimeter hits (simulated hits in this case), reconstruction of tracks in the tracking system using pattern recognition algorithms, and particle flow objects reconstruction with PandoraPFA. For the CLIC detectors, there is an extra step of $\gamma\gamma \rightarrow \texttt{hadrons}$ background suppression. We will discuss PandoraPFA citeThomson:2009rp,Marshall:2012ry,Marshall:2015rfa in details, where a lot work goes into, and the background suppression due to its relevance in later analysis.

3.2 Pandora

Inputs of PandoraPFA are digitised calorimeter hits and reconstructed tracks.

3.2.1 Track selection

Tracks are selected based on their topological properties, how likely they are from physical processes, and whether they are consistent with tracker resolution. Only selected tracks will be used for the subsequent reconstruction.

Using a helical fit of last 50 reconstructed hits, tracks are projected to the front of the ECal.

3.2.2 Calorimeter selection

Calorimeter hits are selected based on a series of criterion. The selected hits need to have energies above the threshold, using the conversion of a minimum ionising particle (MIP) equivalent.

Isolated hits, often originated from low energy neutrons in a hadronic shower, are difficult to associate to the correct hadronic shower. They are identified and not used in the clustering. But their energy is added in the particle flow object (PFO) creation step.

3.2.3 Clustering

The main clustering scheme used in PandoraPFA is cone clustering, for grouping calorimeter hits. Cone clustering has a specified opening angle of the seed hit. Because the direction of particle flows is largely unchanged from the originated particle, whether it is a electromagnetic shower, QCD radiation or hadronisation, these cone clusters have similar direction and energy to the originated particle.

Typically a high energy calorimeter hit will be chosen as a "seed". A cone with a specified opening angle and depth will be formed around the seed. The four-momentum of calorimeter hits sum to the cone's four-momentum. PandoraPFA will start to use projection of tracks at the ECal as seeds. When all tracks are used, the remaining hits will be used as seeds.

These cone clustering algorithms are widely used in the calorimeter in PandoraPFA, and they produce basic working objects, Clusters.

There are two standalone particle identification algorithms in PandoraPFA, muon identification and photon identification. Identified muons and photons will not participate

in the clustering and re-clustering stages. Both algorithms aim to improve the clustering and the re-clustering. The photon identification and related algorithm will be discussed in details in chapter **??**.

3.2.4 Clusters Merging

Initial clustering scheme is aggressive at splitting clusters. The next step is to merge clusters base on clear topological signatures. For clusters associated to tracks (charged clusters) and clusters not associated to tracks(neutral clusters), track like segments in the calorimeter are identified.

These merging signatures include combining track segments, connecting tack segments with gaps, connecting track segment to a hadronic shower, and merging clusters when they are within close proximity.

3.2.5 Re-clustering

The clustering and cluster merging scheme work well for low energy (less than $50\,\mathrm{GeV}$) jet. For a high energy jet, particles and the subsequent hadronic showers are more boosted and more likely to overlap each other. herefore, it is important to re-cluster base on the compatibility of the cluster energy and the associated track momentum. A cluster may be split into two, or two clusters maybe be re-clustered based on the track-cluster association. The re-clustering algorithm is applied iteratively to find a more correct clustering of calorimeter hits.

3.2.6 Photon identification

The neutral clusters are tested against an expected photon electromagnetic shower profile. The longitudinal shower profile for a photon cluster is required to be similar to a expected electromagnetic shower profile, with the discrepancy being smaller than a threshold.

3.2.7 Fragment removal

The late stage of the reconstruction will focus on merging low energy clusters, especially non-photon neutral clusters. These neutral clusters are likely to be fragments of charged

clusters, instead of being a physical particle. The merging criterion are mostly based on the proximity and the energy comparison.

One algorithm will attempt to split up photon clusters, where each is originated from two close by photons. This will be described in details in chapter **??**.

3.2.8 Particle Flow Object Creation

Particle Flow Objects (PFOs) are created at the last step. Tracks are associated to the clusters based on the proximity. Simple but effective particle identification for electrons, muons are applied. Photon identifications have been applied at various stages of the reconstruction.

PFOs are the output of the PandoraPFA reconstruction. The four-momentum of these PFOs are used heavily for the downstream analysis. The electron, muon and photon identification are also used in physics analysis, such as one described in chapter **??**.

3.3 Suppression of $\gamma\gamma \rightarrow \mathtt{hadrons}$ backgrounds

For the CLIC, as discussed in section **??**, significant $\gamma\gamma \rightarrow \mathtt{hadrons}$ background is present. It is crucial to remove the beam induced background as they don't represent the underlying physics process.

Two Marlin process has been developed to suppress these background, a track selector and a PFO selector [2].

The track selector aims to remove poor quality and fake tracks. It places simple quality cut and a simple time of arrival cut. If the arrival time of the track at the front of the ECal, using the helical fit, differs more than $50\,\mathrm{ns}$ from using a straight line fit, the track will be rejected.

The PFO selector utilise the high spatial resolution from the high granular calorimeter. PFOs from $\gamma\gamma \rightarrow \mathtt{hadrons}$ often have low $\mathtt{p_T}$ and have a range of time. PFOs from physics processes have a range of $\mathtt{p_T}$, and have time close to the brunch crossing time. These two distinctive features allow $\gamma\gamma \rightarrow \mathtt{hadrons}$ background to be separated. The optimal suppression uses different $\mathtt{p_T}$ and time cuts for the central part of the detector, and for the forward part of the detector, and uses different cuts for photons, neutral PFOs

and charged PFOs. Three configurations of these cuts are developed, namely "loose", "normal", and "tight" selections. As the name suggested, "loose" selection corresponds to a looser cut of p_T and time. The optimal configuration depends on the $\sqrt{s}$ of the collision, and the physics process to study.

The background suppression is used in analysis described in chapter 6

Chapter 4

Photon Reconstruction

"Photons have mass? I didnąŕt even know they were Catholic."
   — Woody Allen

Chapter 5

Tau Lepton Final State Separation

"MVA: Turn numbers into gold."
  — TMVA

5.1 Introduction

Why study tau

Tau lepton has been examined closely in the past. The decay and the spin of the decay product were direct tests to the standard model. The spin of the decay product, using a Higgs decaying to tau tau channel, allows one to determine the spin of the higgs. Also, as tau is short-lived, only its decay products can be detected and reconstructed in the detector. Therefore, the ability to reconstruct and separate different tau decay modes is benchmark of detector performances.

5.2 Analysis

This chapter will describe a tau final decay separation study. The processor developed for the study is used to test different detector models, as a proof-of-principle of detector optimisation using tau decay separation. Lastly, the spin of the higgs was studied using one higgs decaying to two tau tau channel.

5.3 Simulation and Reconstruction

$e^-e^+ \to \tau^-\tau^+$ channel is used for the tau decay mode separation study. Generator software WHIZARD 1.95 [3] is used to generated simulated Monte Carlo (MC) samples. Hadronisation is described with PYTHIA 6.4 [4], which is tuned to the LEP results []. The spin effect of tau lepton decay is described by TAUOLA [5].

Final state radiation (FSR) was simulated. The initial state radiation (ISR) and the beam induced background were not simulated.

Events were simulated with the CLIC_ILD detector concept, using software with MOKKA [6], based on the GEANT 4 package [7]. Events were reconstructed with ilcsoft version v01-17-07 [8] and PandoraPFA version v02-02-00 [9], where the photon reconstruction is described in [10].

5.4 Generator level cut

To study the difference between different tau decay modes, clear topological difference is required. Therefore, events was considered if the event passes a set of cuts at generator level, listed here

- the final state photons not converting to electron pair in the tracker,

- the tau leptons decaying in the barrel and the end cap regions, which are defined as polar angle between 0.3 to 0.6 rad and 0.8 to 1.57 rad, and

- the visible energy of the tau lepton decay products more than 5 GeV, where the visible energy of the tau lepton decay is defined as the energy of the tau minus the energy of the tau neutrino.

The angular requirement is due to the gap region between the barrel and and the end cap of calorimeters, which degrades the PFO resolution significantly.

Around two million events were simulated for this study.

Table 5.1: Branching fractions of the seven τ^- decays in this study, taken from [11]. τ^+ decays similarly to τ^-.

Decay mode	Branching fraction / %
$e^- \, \overline{\nu}_e \, \nu_\tau$	17.83 ± 0.04
$\mu^- \, \overline{\nu}_\mu \, \nu_\tau$	17.41 ± 0.04
$\pi^- \, \nu_\tau$	10.83 ± 0.06
$\rho(\pi^- \pi^0)_{770}\nu_\tau$	25.52 ± 0.09
$a_1(\pi^- \pi^0 \pi^0)_{1260}\nu_\tau$	9.30 ± 0.11
$a_1(\pi^- \pi^- \pi^+)_{1260}\nu_\tau$	8.99 ± 0.06
$\pi^- \, \pi^- \, \pi^+ \, \pi^0 \, \nu_\tau$	2.70 ± 0.08

5.5 Decay modes

Seven decay final states of the tau lepton shown in table 5.1 were studied, covering 92.58 % of all tau decays. The decay modes not covered have branching fractions lower than 1% each. These final states can be classified into three categories: leptonic decays, one-prong with photons and three-prong with photons.

The channel $e^- e^+ \to \tau^- \tau^+$ contains two τ travelling back-to-back. As we are interested in single τ decay, the detector fiducial space is divided into two halves, each corresponding to one τ. The division was performed with the thrust axis, where defined as $T = \max_{\hat{n}} \frac{\sum_i |p_i \cdot \hat{n}|}{\sum_i |p_i|}$. p_i is the momentum three-vector of a Particle Flow Object (PFO), $\hat{n}$ is the thrust axis, a unit 3-vector that maximise the thrust, T. PFOs were separated into two collections based on the sign of the dot product between the momentum and the thrust axis vector.

Chapter 6

Double Higgs Bosons Production Analysis

"Two is better than one"
    — Sir Steve Orange, 1785–1854

6.1 Analysis Straggly Overview

Proof-of-principle study was perform at CLIC using CLIC_ILD detector model for $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$. Simulated samples, including those containing double higgs production were used. Signal events, events with double higgs production, were selected via a set of carefully designed and complicated methods. g_{HHH} and g_{WWHH} are extracted simultaneously with template fitting with modified couplings samples.

6.2 Monte Carlo Sample Generation

Single channel is defined as $\text{e}^-\text{e}^+ \to \text{HH}\nu\overline{\nu}$. It is divided into sub-channel $\text{HH} \to \text{b}\overline{\text{b}}\text{W}^+\text{W}^-$ and $\text{HH} \to \text{b}\overline{\text{b}}\text{b}\overline{\text{b}}$ to allow closer examination and an improvement of signal selection when combined. In particular, I studied $\text{HH} \to \text{b}\overline{\text{b}}\text{W}^+\text{W}^-$ sub-channel.

Selected background samples, including processes initiated by photons, are considered in the analysis and listed in Table **??**. These background were expected share similar

topologies with the signal process. When describing a multi-quark final state, it is referring to all final states of the same number of quarks, including final states with possible additional neutrinos and or leptons. A multi-quark final state does not include higgs production, unless explicitly stated.

The usual two-quark and four-quark final states were considered. Since the significant presence of beamstrahlung, where photon produced due to the high electric field generated by the colliding beams, processes initiated by photons are also included.

Processes invloving real photons from beamstrahlung (BS) and "quasi-real" photons are generated separately. For the "quasi-real" photon initiated processes, the Equivalent Photon Appproximation (EPA) has been used.

Photon-electron/photon-photon interactions with four-quark final states were considered. Photon-electron interaction with two-quark final state, one Higgs, and one neutrino is considered. Photon-electron interaction with two-quark final state, one Higgs, and one lepton is not considered due to its negligible cross section.

Single higgs productions are not considered because topologies are very different to the single process. Six-quark final states were not considered due to computational limitation.

For processes involving Higgs production explicitly, simulated Higgs mass is 126 GeV. As multi-quark final state background samples could, in principle, contain double higgs production, they are generated with a Higgs mass of 14 TeV. This will produce negligible double higgs production cross section.

All samples are generated with WHIZARD 1.95 [], taking into account the expected CLIC luminosity spectrum. PYTHIA 6.4 [] tuned on LEP data [] is used to describe fragmentation, hadronisation processes, and Higgs decays. TAUOLA [] is used for τ lepton decays.

Simulation

For most background processes, events are simulated when invariant mass of quarks are above 50 GeV. For electron-photon interaction with four quarks and a neutrino final state, events are simulated when invariant mass of quarks are above 120 GeV. These limits are necessary to generate a large amount of background samples in a feasible time, without losing much signal samples.

Finally, the main beam induced background $\gamma\gamma \rightarrow \texttt{hadrons}$ is simulated and overlayed [] to all samples according to the integration time of each subdetector.

6.3 Physics object and event reconstruction

Simulation is performed by MOKKA, interfacing GEANT 4. The reconstruction is done via Marlin in iLCSoft v01-16. Separate software package (processor) exists for identification of electrons, muons, taus, and jet reconstruction. New processors have been developed and existing processors have been optimised for a compromise of signal selection and background rejection. The latest function flavour tagging processor exist in iLCSoft v01-16, which prevented the usage of more recent iLCSoft.

For my signal channel, $HH \rightarrow b\bar{b}W^+W^-$, there is no lepton in the final state. Hence a effective lepton identifier would improve the signal identification. Processors are wither developed or optimised with samples at $\sqrt{s} = 1.4\,\mathrm{TeV}$, and checked against samples at $\sqrt{s} = 3\,\mathrm{TeV}$. Because the expected signal significance would be low, the processors are optimised to reject more background at the cost of losing a bit more signals, to increase the signal significance. It was found that the same set of parameters work well under $\sqrt{s} = 1.4\,\mathrm{TeV}$ and $3\,\mathrm{TeV}$.

6.3.1 Electron and muon identification

IsolatedLeptonFinderProcessor

In Marlin package, IsolatedLeptonFinderProcessor has been used. The optimal parameters ware chosen in collaboration and tested. The particle is identified as an isolated light lepton if it passes a chain of cuts.

A charge track is considered if it has more than $15\,\mathrm{GeV}$ energy. An electron is identified if the energy in the ECal is over 90% of the total calorimetric energy. A muon is identified if the energy in the ECal is between 5% and 25% of the total calorimetric energy. Furthermore, only primary track is selected, which requires the Euclidean distance in the x-y plane, the in z direction, and in the x-y-z three dimensional space of the track starting point to the impact point to be less than $0.02\,\mathrm{mm}$, 0.03mm, and $0.04\,\mathrm{mm}$, respectively.

Channel	$\sigma(\sqrt{s} = 1.4\,\text{TeV})$ / fb	$\sigma(\sqrt{s} = 3\,\text{TeV})$ / fb
$e^- e^+ \to HH\nu\bar{\nu}$	0.588	0.149
$e^- e^+ \to q_l q_l H\nu\bar{\nu}$	0.86	1.78
$e^- e^+ \to c\bar{c}H\nu\bar{\nu}$	0.36	1.12
$e^- e^+ \to b\bar{b}H\nu\bar{\nu}$	0.31	1.91
$e^- e^+ \to qqqq$	1245.1	546.5
$e^- e^+ \to qqqq\ell\ell$	62.1	169.3
$e^- e^+ \to qqqq\ell\nu$	110.4	106.6
$e^- e^+ \to qqqq\nu\bar{\nu}$	23.2	71.5
$e^- e^+ \to qq$	4009.5	2948.9
$e^- e^+ \to qq\ell\nu$	4309.7	5561.1
$e^- e^+ \to qq\ell\ell$	2725.8	3319.6
$e^- e^+ \to qq\nu\nu$	787.7	1317.5
$e^-\gamma(\text{BS}) \to e^- qqqq$	1160.7	1268.7
$e^+\gamma(\text{BS}) \to e^+ qqqq$	1156.3	1267.6
$e^-\gamma(\text{EPA}) \to e^- qqqq$	287.1	287.9
$e^+\gamma(\text{EPA}) \to e^+ qqqq$	286.9	287.8
$e^-\gamma(\text{BS}) \to \nu qqqq$	136.9	262.5
$e^+\gamma(\text{BS}) \to \bar{\nu} qqqq$	136.4	262.3
$e^-\gamma(\text{EPA}) \to \nu qqqq$	32.6	54.2
$e^+\gamma(\text{EPA}) \to \bar{\nu} qqqq$	32.6	54.2
$e^-\gamma(\text{BS}) \to qqH\nu\nu$	15.8	58.6
$e^+\gamma(\text{BS}) \to qqH\nu\nu$	15.7	58.5
$e^-\gamma(\text{EPA}) \to qqH\nu\nu$	3.39	11.7
$e^+\gamma(\text{EPA}) \to qqH\nu\nu$	3.39	11.7
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	21406.2	13050.3
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	4018.7	2420.6
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	4034.8	2423.1
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	753.0	402.7

Table 6.1: List of signal and background samples with the corresponding cross sections at $\sqrt{s} = 3\,\text{TeV}$ and $\sqrt{s} = 1.4\,\text{TeV}$. q can u, d, s, b or t. Unless specified, q, ℓ and ν represent particles and its corresponding anti-particles. γ (BS) represents a real photon from beamstrahlung (BS). γ (EPA) represents a "quasi-real" photon, simulated with the Equivalent Photon Approximation. For processes involving Higgs production explicitly, simulated Higgs mass is 126 GeV. Otherwise, Higgs mass is set to 14 TeV. Simulated W has invariant mass of 80.385 GeV.

The isolation criteria states that

$$E_{cone}^2 \leqslant 5.7 \times E_l - 50 \tag{6.1}$$

where, E_{cone} is the total energy of PFOs within an opening angle of $\cos^{-1}(0.995)$ of the light lepton, and E_l is the energy of the light lepton.

BonoLeptonFinderProcessor

The IsolatedLeptonFinderProcessor is rather conservative. I developed a new more aggressive light lepton selection processor, BonoLeptonFinderProcessor, that utilises calorimetric information provided by PandoraPFA.

The processor uses two chains of cuts.

First chain uses the particle ID information from PandoraPFA. A electron is identified if it is a "PandoraPFA" electron and the energy in the ECal is over 95% of the total calorimetric energy. A muon is identified if it is a "PandoraPFA" muon. Primary track selection states the Euclidean distance in the x-y-z three dimensional space of the track starting point to the impact point to be less than $0.015\,\text{mm}$, and the PFO energy is more than $10\,\text{GeV}$. The light lepton either satisfy the high p_T requirement of at least $40\,\text{GeV}$, or the isolation criteria,

$$E_l \geqslant 23 \times \sqrt{E_{cone}} + 5 \tag{6.2}$$

where E_{cone} and E_l have the same definition as in the IsolatedLeptonFinderProcessor.

Second chain of cuts is similar to the IsolatedLeptonFinderProcessor. An electron is identified if the energy in the ECal is over 95% of the total calorimetric energy. A muon is identified if the energy in the ECal is between 5% and 20% of the total calorimetric energy. Primary track selection states the Euclidean distance in the x-y-z three dimensional space of the track starting point to the impact point to be less than $0.5\,\text{mm}$, and the PFO energy is more than $10\,\text{GeV}$. The light lepton either satisfy the high p_T requirement of at least $40\,\text{GeV}$, or the isolation criteria,

$$E_l \geqslant 28 \times \sqrt{E_{cone}} + 30 \tag{6.3}$$

where, E_{cone} is the total energy of PFOs within an opening angle of $\cos^{-1}(0.99)$ of the light lepton, and E_l is the energy of the light lepton.

Comparison: IsolatedLeptonFinderProcessor v.s. BonoLeptonFinderProcessor

Two processors share similar criterion for light lepton identification. The main difference is that the BonoLeptonFinderProcessor allows high p_T light lepton to be identified in a potential non-isolated environment, which leads to the more aggressiveness of the BonoLeptonFinderProcessor. The performance of two processors on the signal and selected background samples is shown in table 6.2

6.3.2 Tau identification

TauFinderProcessor

With a decay length of $87\mu m$, tau leptons decay before reaching the detector and can only be identified through the reconstruction of their decay products. The leptonic decay of tau can be identified using the two isolated lepton finder processor. Therefore tau identification will focus on the hadronic decay.

TauFinderProcessor [12], an existing processor Marlin package, has been tuned in collaboration and tested. The a collection of tau decay productions are identified they pass a chain of cuts.

Particles are not considered if p_T is less than $1\,\text{GeV}$ or $|\cos(\theta_z)|$ is more than $1.1\,\text{rad}$, as they are more likely from beam induced background. A seed is considered if a charged particle has p_T more than $10\,\text{GeV}$. A search cone of opening angle $0.03\,\text{rad}$ is then formed. The search cone is rejected if it has more than 3 charged particles, more than 10 particles or its invariant mass more than $2\,\text{GeV}$. An isolation cone is formed with opening angle between 0.03 and $0.33\,\text{rad}$ of the seed. The seed is rejected if there are more than $3\,\text{GeV}$ in the isolation cone.

BonoTauFinderProcessor

The TauFinderProcessor's performance is decent, but there is room for improvement. I developed a new more aggressive tau lepton selection processor, BonoTauFinderProcessor, that utilises calorimetric information provided by PandoraPFA.

Similar to the previous processor, PFOs with p_T less than $1\,\mathrm{GeV}$ are rejected. A tau seed is defined as a charged particle with p_T at least $5\,\mathrm{GeV}$. The search cone has an opening angle of $\cos^{-1}(0.999)$. Particles are iteratively added to the search cone according to the size of the opening angle to the seed. A temporary search cone is then considered if it has one or three charged particles, and the invariant mass is less than $3\,\mathrm{GeV}$. The search cone needs to satisfy one of isolation criterion.

1. No particle in the large isolation cone, and p_T of search cone at least $10\,\mathrm{GeV}$,

2. One charged particle in the search cone, one particle in the large isolation cone, and r_0 larger than $0.01\,\mathrm{mm}$,

3. Three charged particle in the search cone, one particle in the large isolation cone, p_T of search cone at least $10\,\mathrm{GeV}$, and search cone opening angle less than $\cos^{-1}(0.9995)$,

4. One charged particle in the search cone, no particle in the small isolation cone, r_0 larger than $0.01\,\mathrm{mm}$, and p_T of search cone at least $10\,\mathrm{GeV}$,

5. Three charged particle in the search cone, no particle in the small isolation cone, p_T of search cone at least $10\,\mathrm{GeV}$, and search cone opening angle less than $\cos^{-1}(0.9995)$,

where large and small isolation cone are defined as opening angle of $\cos^{-1}(0.95)$, and $\cos^{-1}(0.99)$ respectively. If there are multiple temporary search cone of a same seed passing the isolation criteria, the cone with smallest opening angle is chosen for output.

Comparison: TauFinderProcessor v.s. BonoTauFinderProcessor

Two processors share similar size of search cone and isolation cone. The BonoTauFinder-Processor has looser cut on minimum p_T and invariant, but stricter isolation criterion. This leads to a more aggressive tau finder. The performance of two processors on the signal and selected background samples is shown in table 6.2

Selection / Efficiency (1.4 TeV)	Signal	$qqqq\ell\nu$
IsolatedLeptonFinderProcessor	99.3%	50.3%
BonoLeptonFinderProcessor	99.1%	39.9%
TauFinderProcessor	97.5%	52.3%
BonoTauFinderProcessor	89.7%	38.5%
ForwardFinderProcessor	98.9%	95.1%
Combined	86.6%	16.8%
Processor / Efficiency (3 TeV)	Signal	$qqqq\ell\nu$
IsolatedLeptonFinderProcessor	99.5%	66.8%
BonoLeptonFinderProcessor	99.0%	52.5%
TauFinderProcessor	97.7%	79.5%
BonoTauFinderProcessor	86.3%	60.3%
ForwardFinderProcessor	95.9%	80.7%
Combined	81.0%	23.3%

Table 6.2: isolated lepton finder processors performance on the signal and selected background samples.

6.3.3 Very forward electron identification

Certain background channels, for example photon-electron interactions, contain electrons in the very forward part of the detector, namely LCal and BCal. These forward calorimeters were not simulated due to computational limitation. Most particle in these detector would be very forward particles from beam induced background. However, previous study has shown [] that high energy electrons can be identified with high efficiency. Due to the lack of tracking in these region, electrons and photons would have the same electromagnetic shower profile, with the given calorimeter resolution. MC photons and electrons are checked if they fall in the LCal or the BCal, and checked against the known detection efficiency.

Beam Calorimeter acceptance is defined as $|\cos(\theta_Z)|$ is between 0.01 and 0.04 rad and length in z direction is between 3181 and 3441 mm. Luminosity Calorimeter acceptance is defined as $|\cos(\theta_Z)|$ is between 0.038 and 0.11 rad and length in z direction is between 2539 and 2714 mm. For $\sqrt{s} = ($ TeV 3$)$, the BeamCal detection efficiency is provided by a software package []. For $\sqrt{s} = ($ TeV 1.4$)$, the same software for the BeamCal is used, by scaling the energy of the MC particle by a factor of $\frac{3}{1.4}$. For the LumiCal, the

Selection / Efficiency (1.4 TeV)	Signal	$e^-\gamma(BS) \to e^- qqqq$
Combined light lepton finder	87.6%	67.5%
ForwardFinderProcessor	98.9%	53.6%
Combined	86.6%	30.8%
Processor / Efficiency (3 TeV)	Signal	$e^-\gamma(BS) \to e^- qqqq$
Combined light lepton finder	84.4%	72.7%
ForwardFinderProcessor	95.9%	55.4%
Combined	81.0%	33.4%

Table 6.3: Very forward electron and photon finder performance on the signal and selected background samples.

identification efficiency is defined as

$$\varepsilon = \begin{cases} 0, & \text{if } E < 50 \,\text{GeV} \\ 0.99 \times \frac{(\text{erf}(E-100)+1)}{2}, & \text{otherwise} \end{cases} \tag{6.4}$$

where E is the energy of the electron or the photon.

The background rejection is significant, shown in table **??** for the signal and selected background.

6.3.4 Other lepton identification processors

Other isolated lepton selection processors available in Marlin package, including IsolatedLeptonTagging and TauJetClustering, have been tested. The results, after some tuning of parameters, were unsatisfactory. They either performed poorly comparing to the processors above, or became redundant after the processors above. Therefore, these processors were not used in this analysis.

6.4 Jet reconstruction

The signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, is a four-jet final state. A useful technique for the analysis is to reconstruct the four-jet final state using jet algorithms. This allows discriminative variables to be calculated.

6.4.1 Jet reconstruction optimisation

Longitudinal invariant, k_t, jet algorithm was chosen for the jet clustering. Due to the presence high level of beam induced background at the CLIC, it has been shown that a jet algorithm designed for hadron colliders are more effective than those traditional designed for the electron-positron collider, such as Durham algorithm. []

The free parameters for k_t algorithm is the R parameter, which controls the fatness of the jet. There is also the choice of the PFO collection, which incorporate different level of time and p_T cuts, to reduce beam induce background. Both parameters are optimised for $\sqrt{s} = 1.4\,\mathrm{TeV}$ and $\sqrt{s} = 3\,\mathrm{TeV}$.

The details of jet algorithm can be found in section **??**.

The R parameter of the k_t jet algorithm, and the collection of the PFOs are chosen to give the best invariant mass resolution. When there are a few suitable candidate, analysis were performed in parallel. Decision were made to give the highest signal significance.

k_t jet algorithm was used as part of the FastJet algorithms available in the Marlin package.

The samples containing the signal, $HH \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}qqqq$, was used for the optimisation of the jet reconstruction. The signal events were chosen using MC truth information.

Jet algorithm was run in exclusive mode, where number of jets is chosen to be six.

For the signal, $HH \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}qqqq$, one Higgs decays to two b quarks, resulting in two jets from hadronisation. Similarly the other Higgs decays to two W bosons, where each W boson decays into two quarks. Therefore, the expected number of jets is six.

Jets produced by the k_t jet algorithm are paired up using MC truth information, to the corresponding Higgs and W boson. Four invariant mass distributions are obtained: two Higgs masses, $m_{H_{bb}}$, $m_{H_{WW^*}}$, and two W masses m_W, m_{W^*}. W^* indicates the off-mass-shell W boson, because when a Higgs decays into two W bosons, one W is off the mass shell, as the Higgs mass is less than the sum two W masses.

Three mass distributions are worth comparing for different jet reconstruction, namely, $m_{H_{bb}}$, $m_{H_{WW^*}}$, and m_W. The ideal jet reconstruction should produce the a sharp mass peak around the particle's true mass.
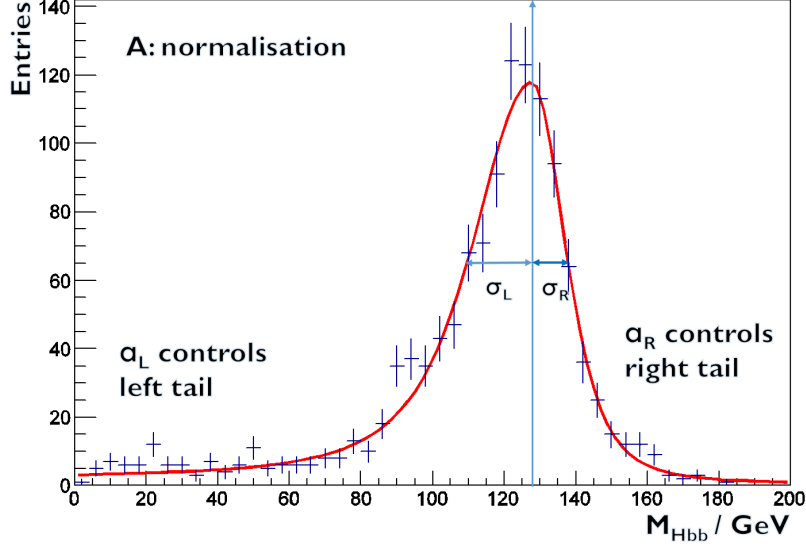
Figure 6.1: A typical example of MC mass fit of $m_{H_{bb}}$ for double higgs analysis. Red line indicates the best fit. Vertical arrow indicates the fitted peak position.

To quantitatively access the mass distribution, a gaussian like fit is performed to extract the position of the peak, and the width of the distribution. The fit has the form:

$$f(m) = Ae^{-\frac{(m-\mu)^2}{g}} \begin{cases} g = 2\sigma_L + \alpha_L(m-\mu), & \text{if } m < \mu \\ g = 2\sigma_R + \alpha_R(m-\mu), & \text{if } m \geqslant \mu \end{cases} \quad (6.5)$$

The fit represents an asymmetrical gaussian function, where m is binned mass distribution, with 50 bins in range $[0, 200]$ GeV. The fitted mass peak is denoted by μ. σ_L and σ_R allow asymmetrical width of the distribution. α parameter controls the fit of tails. Inspired by the $t\bar{t}$ analysis [], the use of the α parameter allows the fit in the whole mass range, otherwise only the peak of the distribution should be fitted with a gaussian like function. A is the normalistion factor. An example of the fit of $m_{H_{bb}}$ is shown in figure 6.1.

For $\sqrt{s} = 1.4$ TeV, shown in figure 6.2, normal selected PFO with $R = 0.7$ give a good fitted mass for H_{WW^*} and W. The mass is slightly too low for the H_{bb}. figure 6.3 shows the combined relative fitted width for the H_{bb}, H_{WW^*} and W. Normal selected PFO with $R = 0.7$ gives an almost optimal relative width for H_{bb}, while achieving a good balance for H_{WW^*} and W. Therefore, normal selected PFO with $R = 0.7$ is chosen to be the optimal jet reconstruction parameters.
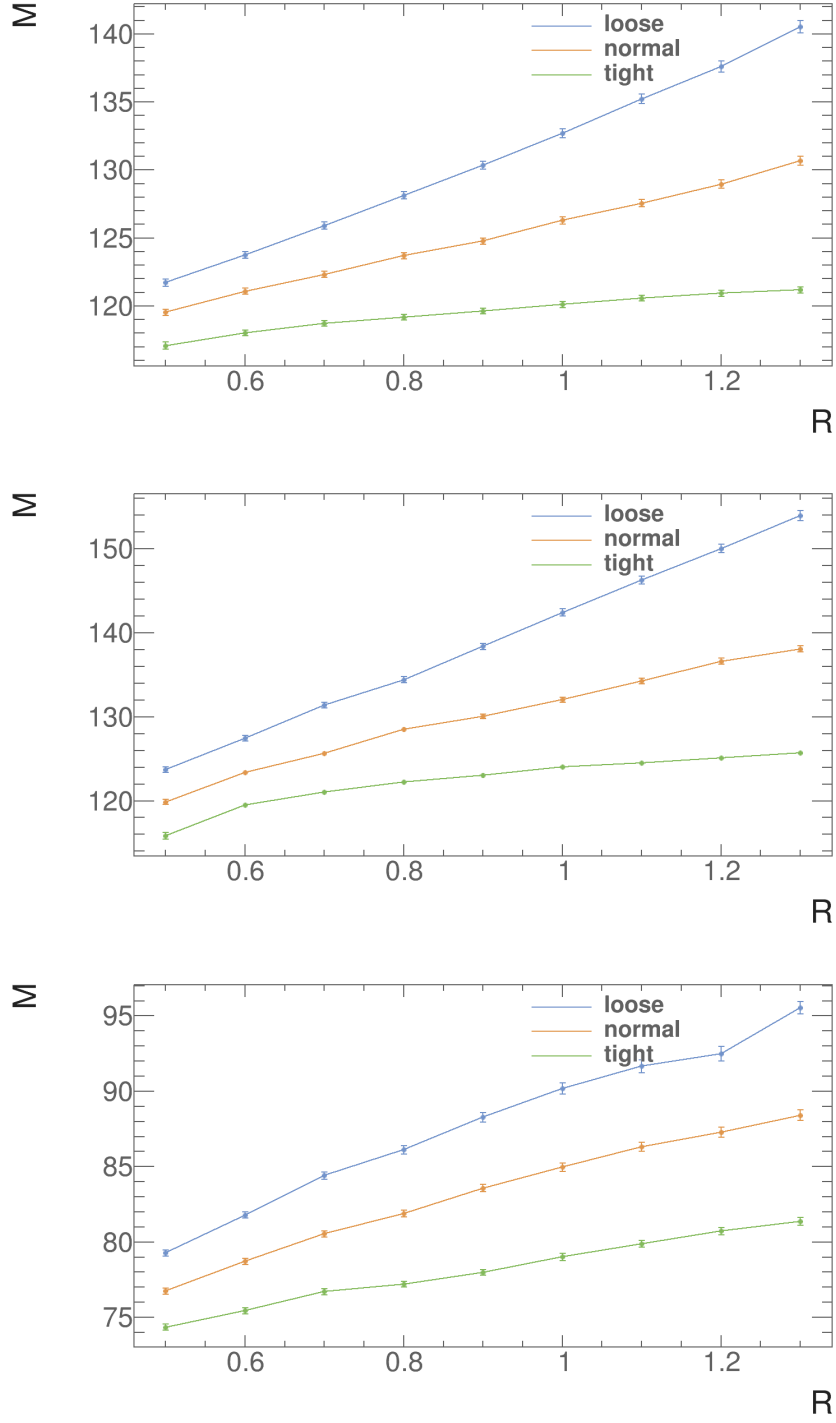
Figure 6.2: Fitted mass and statistical error of H_{bb}, H_{WW^*} and W for $\sqrt{s} = 1.4\,\text{TeV}$, for loose, normal and tight selected PFO against R parameter.
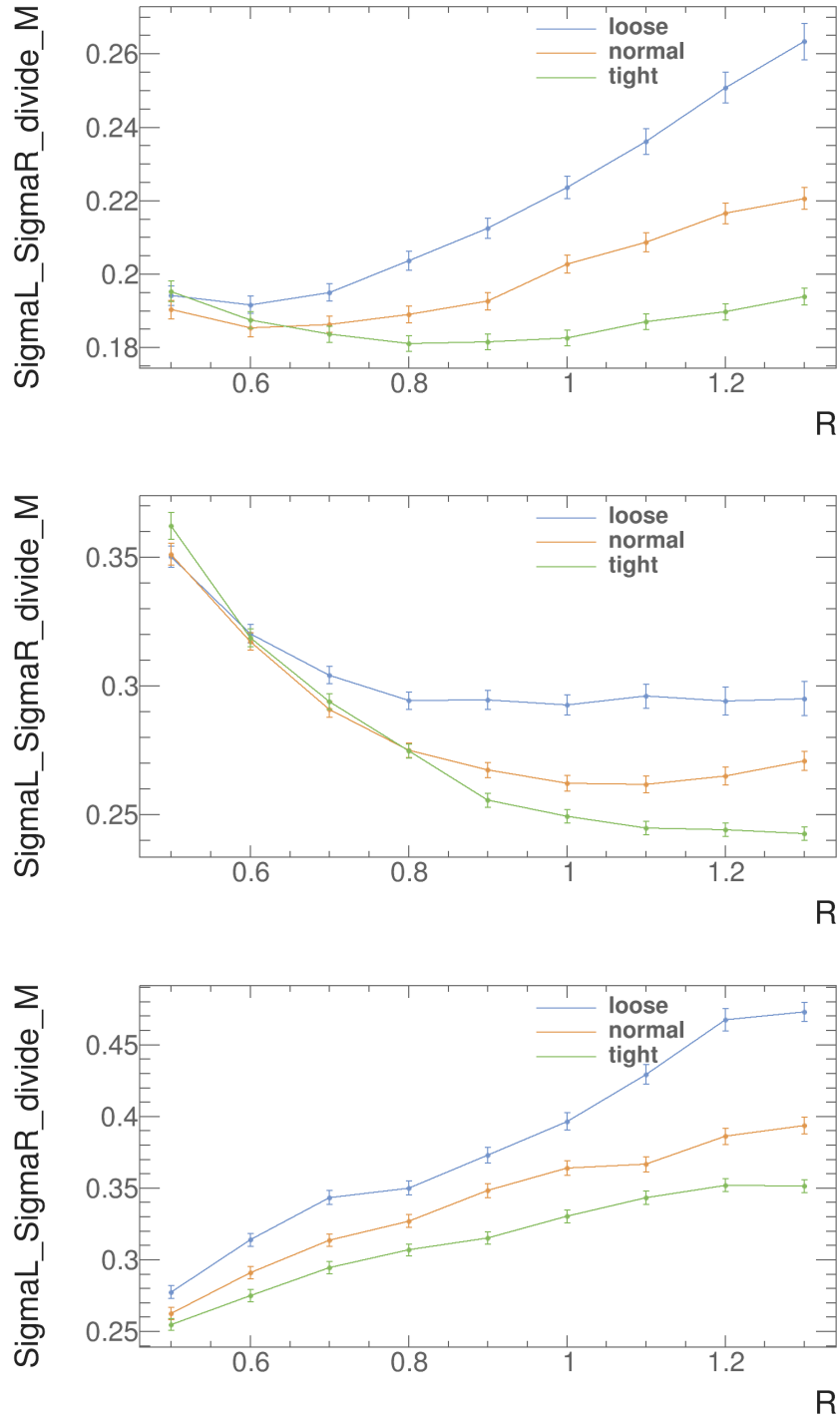
Figure 6.3: Fitted combined width and statistical error of H_{bb}, H_{WW^*} and W for $\sqrt{s} = 1.4\,\text{TeV}$, for loose, normal and tight selected PFO against R parameter.

Jet Parameters	$\sqrt{s} = 1.4\,\mathrm{TeV}$	$\sqrt{s} = \mathrm{S\,TeV}$
$\mu_{H_{bb}}$	$122.3_{\pm0.2}$	$119.1_{\pm0.3}$
$\sigma_{L,H_{bb}}$	$15.2_{\pm0.2}$	$15.0_{\pm0.3}$
$\sigma_{R,H_{bb}}$	$7.55_{\pm0.16}$	$8.4_{\pm0.2}$
$\mu_{H_{WW^*}}$	$125.7_{\pm0.2}$	$123.0_{\pm0.3}$
$\sigma_{L,H_{WW^*}}$	$29.4_{\pm0.3}$	$36.6_{\pm0.6}$
$\sigma_{R,H_{WW^*}}$	$7.18_{\pm0.17}$	$7.4_{\pm0.2}$
μ_W	$80.5_{\pm0.2}$	$78.1_{\pm0.3}$
$\sigma_{L,W}$	$16.2_{\pm0.3}$	$13.1_{\pm0.4}$
$\sigma_{R,W}$	$9.03_{\pm0.16}$	$9.5_{\pm0.2}$

Table 6.4: The extracted fitted parameters of optimal jet reconstructions, normal selected PFO with $R = 0.7$ for $\sqrt{s} = 1.4\,\mathrm{TeV}$ and tight selected PFO with $R = 0.7$ for $\sqrt{s} = 3\,\mathrm{TeV}$.

For $\sqrt{s} = 3\,\mathrm{TeV}$, the choice is a bit more complicated. Shown in figure 6.4, fitted mass for H_{bb} favours normal selected PFO with $R = 0.8$. Fitted mass for H_{WW^*} favours tight selected PFO with $R = 0.9$. Fitted mass for W favours tight selected PFO with $R = 0.8$. Looking at the combined relative fitted width for the H_{bb}, H_{WW^*} and W, shown in figure 6.5, normal selected PFO gives a larger width than tight selected PFO. Within tight selected PFO, small R values provide a shaper width for H_{WW^*} and H_{bb}, but a broader width for W. Therefore, tight selected PFO with $R = 0.7$ and $R = 1$ are both chosen for parallel analysis.

Later it was shown that tight selected PFO with $R = 0.7$ gives a better signal significance. Therefore the optimal choice of jet reconstruction for $\sqrt{s} = 3\,\mathrm{TeV}$ is tight selected PFO with $R = 0.7$.

The extracted fitted parameters of optimal jet reconstructions are summarised in table 6.4.

6.4.2 Jet flavour tagging

Two b-jets out of six jets in final states are identified with flavour tagging processors. The processor calculates a set of discriminatively variables for a jet. After training the
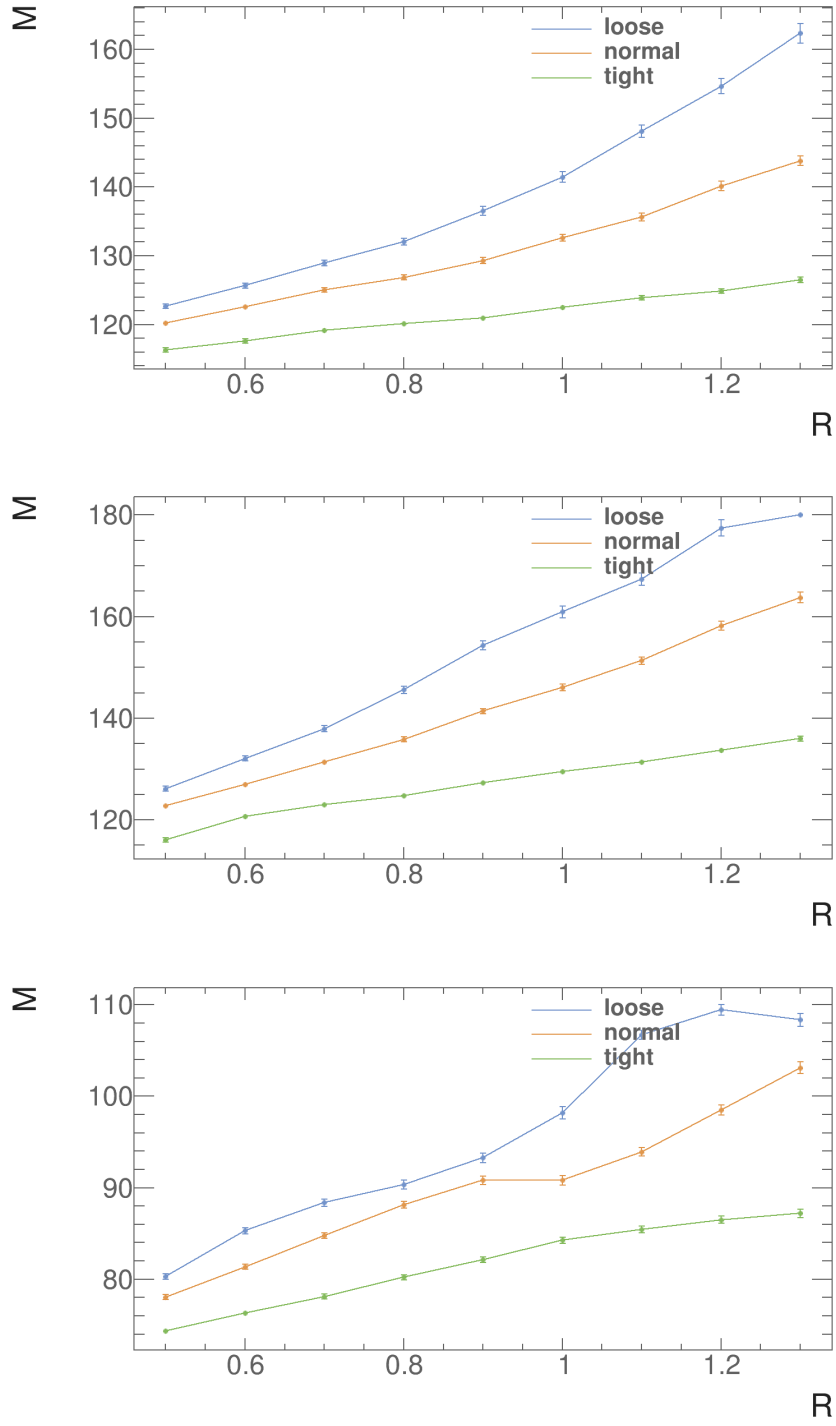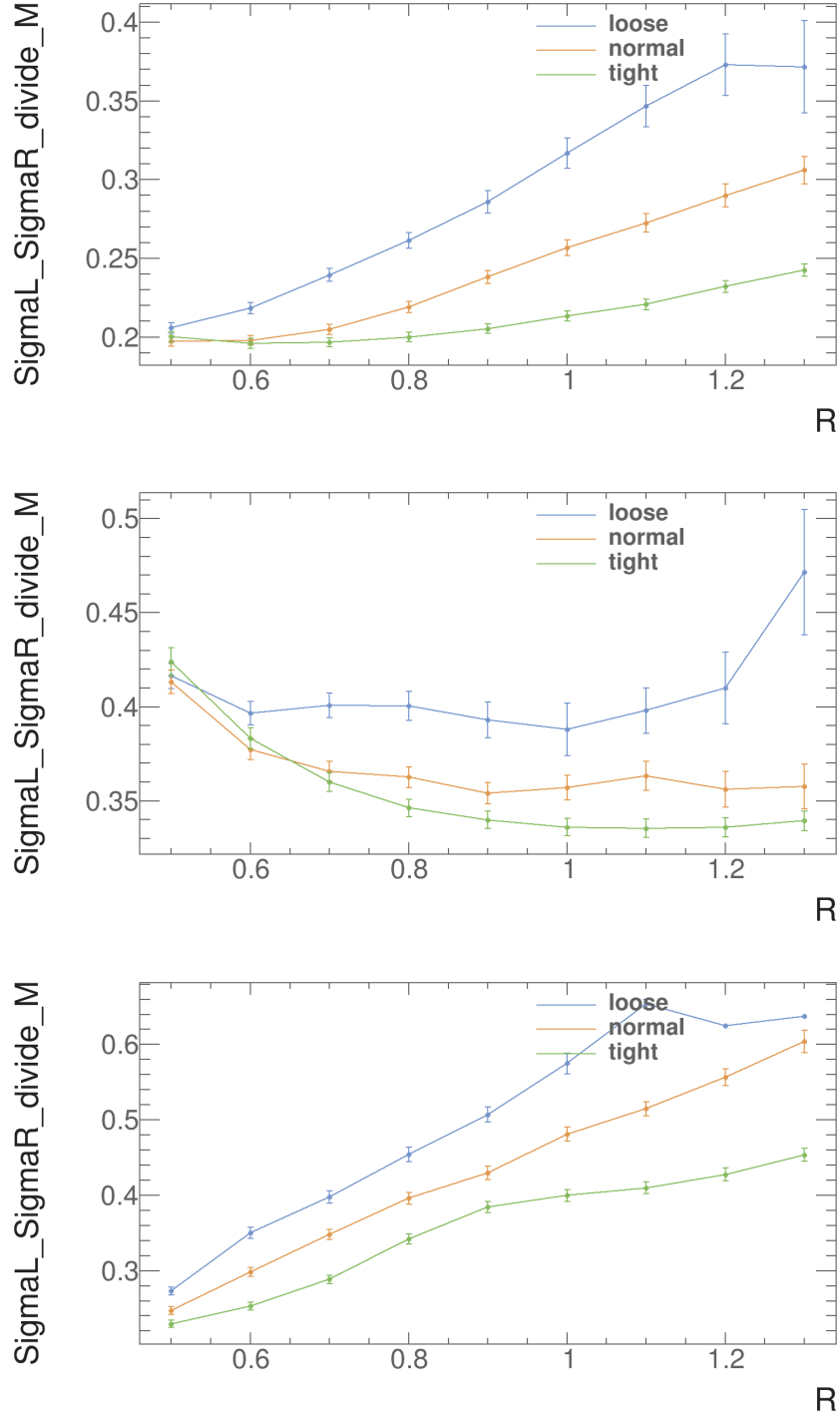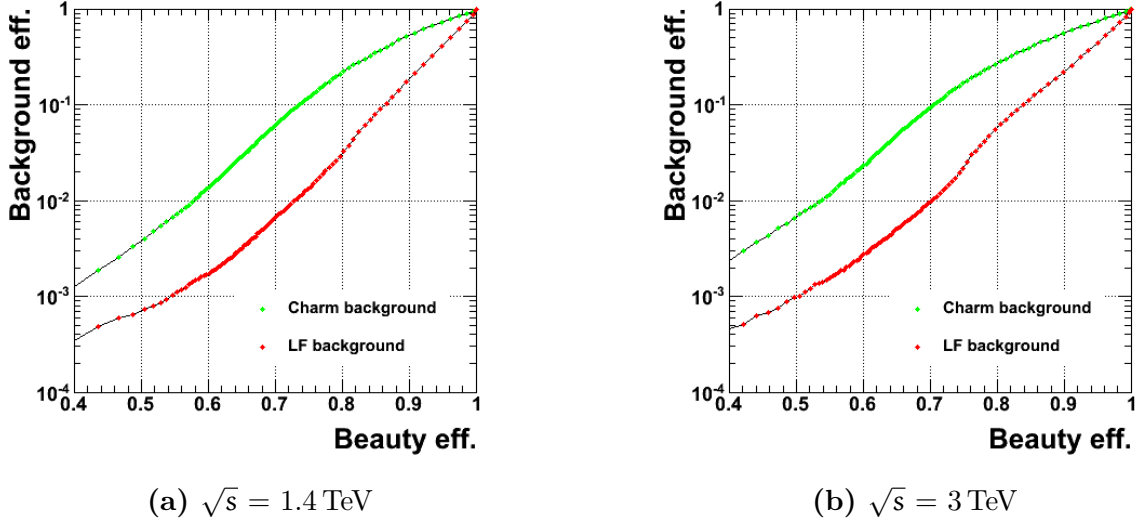
Figure 6.4: Fitted mass and statistical error of H_{bb}, H_{WW^*} and W for $\sqrt{s} = 3\,\text{TeV}$, for loose, normal and tight selected PFO against R parameter.

Figure 6.5: Fitted combined width and statistical error of H_{bb}, H_{WW^*} and W for $\sqrt{s} = 3\,\text{TeV}$, for loose, normal and tight selected PFO against R parameter.

(a) $\sqrt{s} = 1.4\,\mathrm{TeV}$ **(b)** $\sqrt{s} = 3\,\mathrm{TeV}$

Figure 6.6: Performance of b-jet tagging with trainning samples

MVA, the MVA is applied to jets to produce a likelihood for b-jet and c-jet. For details see section **??**.

The existing LCFIPlus processor in Marlin package is used. The training sample of the flavour tagging processor is $\mathrm{e^- e^+ \to Z\nu\overline{\nu}}$, where Z decays to $\mathrm{q_l \overline{q}_l}$, $\mathrm{b\overline{b}}$, or $\mathrm{c\overline{c}}$ at $\sqrt{s} = 1.4\,\mathrm{TeV}$ and $\sqrt{s} = 3\,\mathrm{TeV}$, because they have similar event topology as the signal, and they have only two jets in the final state.

The selection efficiency of b-jets and c-jets with trainning samples are shown in figure **??**. Flavour tagging performs better at low energy. Because at high energy, particles at more collimated and more difficult to separate.

6.4.3 Jet pairing

The jet pairing was performed by seaking combination of jets that are compatible with signal $\mathrm{HH \to b\overline{b}W^+W^-}$.

The actual pairing is done via a minimisation

$$\chi^2 = \left(\frac{m_{ij} - \mu_{H_{bb}}}{\sigma'_{H_{bb}}}\right)^2 + \left(\frac{m_{klmn} - \mu_{H_{WW^*}}}{\sigma'_{H_{WW^*}}}\right)^2 + \left(\frac{m_{kl} - \mu_W}{\sigma'_W}\right)^2, \qquad (6.6)$$

where, $\mu_{H_{bb}}$ and $\sigma'_{H_{bb}}$ are the fitted invariant mass, and the fitted width, respectively. Both are obtained in section 6.4.1. $\sigma'_{H_{bb}}$ is $\sigma_{L,H_{bb}}$ when $m_{ij} < m_{H_{bb}}$, and $\sigma_{R,H_{bb}}$ otherwise. Similarly $\mu_{H_{WW^*}}$ and μ_W are fitted mass, and $\sigma'_{H_{WW^*}}$ and σ'_W are fitted invariant mass, and the fitted width, respectively. Out of the six jets from the jet clustering, indicated by subscript i, j, k, l, m, n, two are used for H_{bb}, two for W and four for H_{WW^*}. The fitted parameters used are listed in table 6.4. Additional requirement is that at least one of two jets forming H_{bb} needs to have a b-jet tag of 0.2 or greater.

With the χ^2, all possible combinations are tested, and the one with smallest χ^2 is chosen.

6.5 Pre-selection

Discriminative variables were calculated. Some are used as to discard background events, whilst hurting the signal events a bit. This allows MVA to concentrate on events where it is difficult to separate in a single parameter space.

6.5.1 Discriminative pre-selection cuts

As discussed before, events with identified leptons are rejected. Jet pairing implies that events with the largest b-jet tag less than 0.2 are rejected.

For $\sqrt{s} = 1.4\,\text{TeV}$, a range of variables were tested and three where chosen as pre-selection cuts.

Event with invariant mass of two higgs less than $150\,\text{GeV}$ is rejected. The cut above $120\,\text{GeV}$ is needed as some background samples were generated only for invariant mass greater than $120\,\text{GeV}$. Shown in table **??** and figure **??**, this cut is effective against samples with two quark final states.

Event with second highest b-jet tag less than 0.2 is rejected. This stricter cut than the jet pairing helps to reduce samples with no b-jets.

Event with p_T of two higgs less than $30\,\text{GeV}$ is rejected. This is extremely effective against samples with no neutrinos in the final state.

For $\sqrt{s} = 3\,\text{TeV}$, event with invariant mass of two higgs less than $150\,\text{GeV}$ is rejected, for the reason similar to $\sqrt{s} = 1.4\,\text{TeV}$.
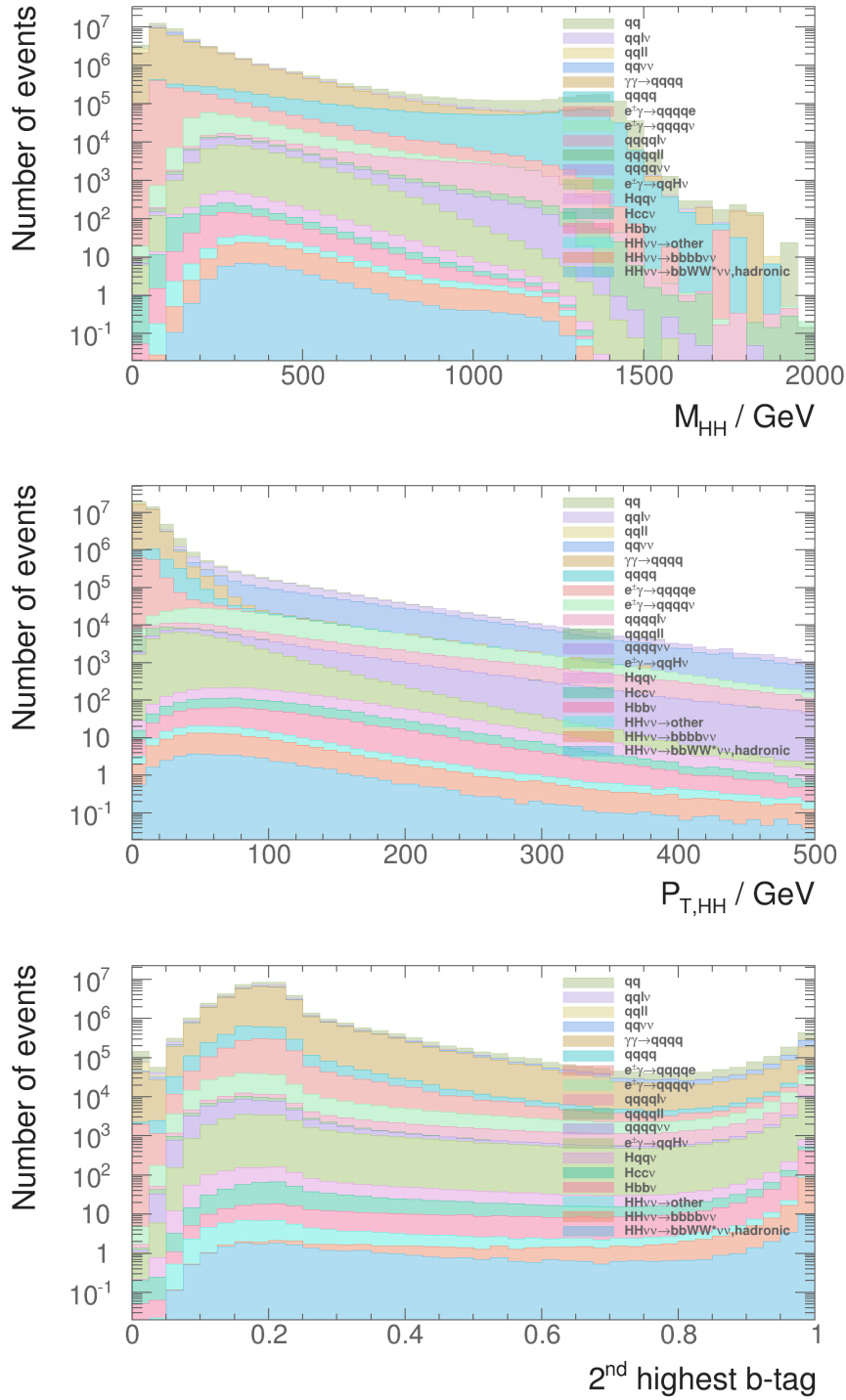
Figure 6.7: Discriminative pre-selection variables for $\sqrt{s} = 1.4\,\text{TeV}$, after rejecting events with identified leptons, and jet pairing
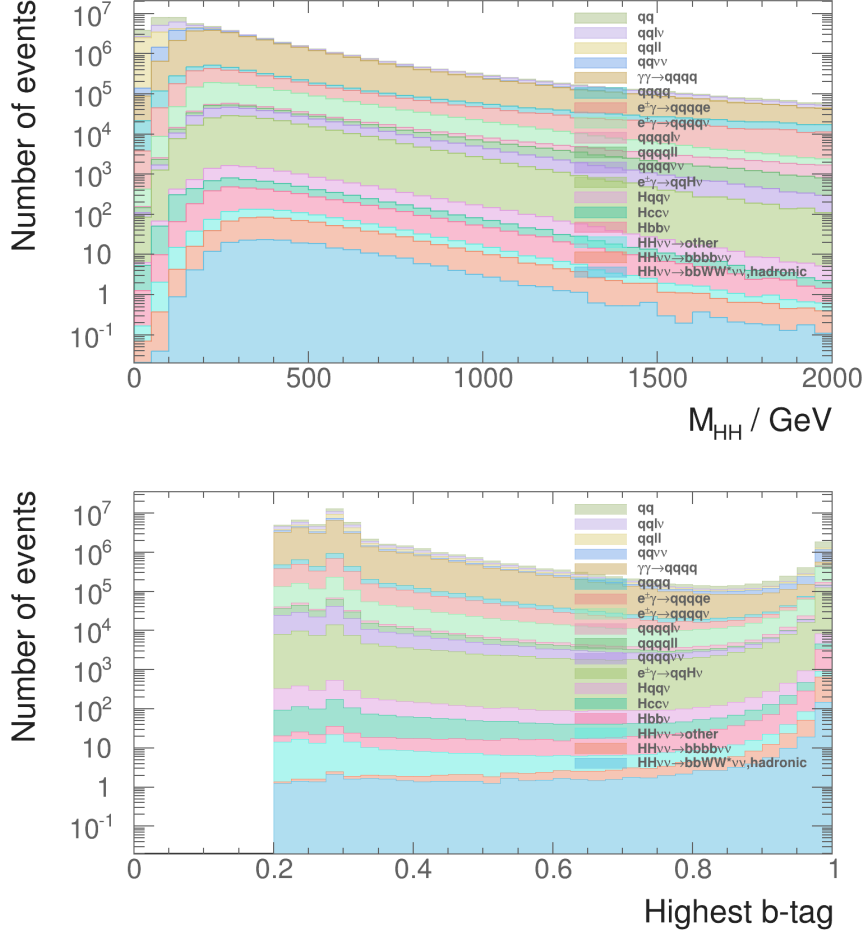
Figure 6.8: Discriminative pre-selection variables for $\sqrt{s} = 3\,\text{TeV}$, after rejecting events with identified leptons, and jet pairing

In addition, event with highest b-jet tag less than 0.7 is rejected. It is found that b-jet tag is less efficient at a higher $\sqrt{s}$. Therefore, a stricter cut at b-jet tag is useful to compensate for the tagging efficiency loss.

These set of cuts are stricter than usual analysis. The cross sections of signal channel for both $\sqrt{s}$ are extremely small, comparing to the background. Hence only the signal events with very clear characteristic topologies would be able to pass the final selection, in order to achieve a decent signal-to-background ratio. Therefore, a strict pre-selection cut would not hurt the final signal selection. On the contrary, final signal selection would benefit from MVA being able to focus the difficult background events, where their topologies are too similar to the signal events to separate in any single parameter space.

Channel / Efficiency $\sqrt{s} = 1.4\,\mathrm{TeV}$	Expected number of events	Lepton ID and jet pairing	$m_{HH} > 150\,\mathrm{GeV}$	$B_2 > 0.2$	$p_T > 30\,\mathrm{GeV}$
$e^-e^+ \to HH\nu\bar{\nu} \to$ $b\bar{b}W^+W^-\nu\bar{\nu}$, hadronic	27.9	85.8%	85.6%	73.7%	66.4%
$e^-e^+ \to HH\nu\bar{\nu} \to$ $b\bar{b}b\bar{b}\nu\bar{\nu}$	67.6	90.8%	90.5%	90.1%	80.6%
$e^-e^+ \to HH\nu\bar{\nu} \to$ other	128.0	36.2%	35.3%	27.7%	24.7%
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	1304.0	60.7%	59.8%	44.9%	42.0%
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	546.1	67.4%	57.7%	46.5%	43.4%
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	463.0	73.9%	72.6%	68.7%	64.2%
$e^-e^+ \to qqqq$	1867650.0	48.8%	46.1%	17.3%	4.7%
$e^-e^+ \to qqqq\ell\ell$	93150.0	5.0%	4.9%	1.5%	0.3%
$e^-e^+ \to qqqq\ell\nu$	165600.0	15.1%	15.1%	12.4%	11.4%
$e^-e^+ \to qqqq\nu\bar{\nu}$	34800.0	50.7%	50.0%	20.1%	18.8%
$e^-e^+ \to qq$	6014250.0	54.5%	17.5%	8.4%	2.2%
$e^-e^+ \to qq\ell\nu$	6464550.0	14.1%	5.3%	2.0%	1.6%
$e^-e^+ \to qq\ell\ell$	4088700.0	13.0%	1.1%	0.6%	0.1%
$e^-e^+ \to qq\nu\nu$	1181550.0	60.1%	12.3%	6.2%	5.8%
$e^-\gamma(BS) \to e^- qqqq$	1305787.5	23.3%	10.6%	4.4%	0.4%
$e^+\gamma(BS) \to e^+ qqqq$	1300837.5	23.4%	10.5%	4.3%	0.4%
$e^-\gamma(EPA) \to e^- qqqq$	430650.0	11.1%	5.4%	2.2%	0.3%
$e^+\gamma(EPA) \to e^+ qqqq$	430350.0	11.1%	5.3%	2.1%	0.3%
$e^-\gamma(BS) \to \nu qqqq$	89775.0	58.3%	56.8%	31.0%	27.7%
$e^+\gamma(BS) \to \bar{\nu} qqqq$	89212.5	57.6%	56.1%	30.3%	27.3%
$e^-\gamma(EPA) \to \nu qqqq$	26100.0	29.6%	28.9%	15.4%	13.9%
$e^+\gamma(EPA) \to \bar{\nu} qqqq$	25950.0	29.2%	28.5%	15.0%	13.7%
$e^-\gamma(BS) \to qqH\nu\nu$	1241.0	61.2%	60.0%	45.4%	34.5%
$e^+\gamma(BS) \to qqH\nu\nu$	1232.8	61.2%	60.0%	45.4%	34.2%
$e^-\gamma(EPA) \to qqH\nu\nu$	354.7	31.9%	31.3%	23.8%	18.2%
$e^+\gamma(EPA) \to qqH\nu\nu$	355.1	32.2%	31.7%	23.9%	18.7%
$\gamma(BS)\gamma(BS) \to qqqq$	2054951.5	56.3%	23.9%	9.6%	0.3%
$\gamma(BS)\gamma(EPA) \to qqqq$	4521037.5	33.6%	14.2%	5.7%	0.4%
$\gamma(EPA)\gamma(BS) \to qqqq$	4539150.0	33.7%	14.2%	5.7%	0.4%
$\gamma(EPA)\gamma(EPA) \to qqqq$	1129500.0	21.1%	9.1%	3.7%	0.4%

Table 6.5: List of signal and background samples with the corresponding expected number at $\sqrt{s} = 1.4\,\mathrm{TeV}$, assuming a luminosity of $1500\,\mathrm{fb}^{-1}$. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.

Channel / Efficiency $\sqrt{s} = 3\,\text{TeV}$	Expected number of events	Lepton ID and jet pairing	$m_{HH} > 150\text{GeV}$	$B_1 > 0.7$
$e^-e^+ \to HH\nu\bar{\nu} \to$ $b\bar{b}W^+W^-\nu\bar{\nu}$, hadronic	146.0	80.2%	79.9%	69.7%
$e^-e^+ \to HH\nu\bar{\nu} \to$ $b\bar{b}b\bar{b}\nu\bar{\nu}$	355.0	83.4%	82.9%	81.2%
$e^-e^+ \to HH\nu\bar{\nu} \to$ other	675.0	36.7%	35.8%	25.2%
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	6115.4	59.5%	58.5%	40.4%
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	2249.9	64.8%	58.4%	39.3%
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	2197.7	69.7%	68.4%	64.2%
$e^-e^+ \to qqqq$	1093000.0	48.5%	39.7%	3.0%
$e^-e^+ \to qqqq\ell\ell$	338600.0	14.7%	14.2%	0.7%
$e^-e^+ \to qqqq\ell\nu$	213200.0	19.7%	19.4%	10.0%
$e^-e^+ \to qqqq\nu\bar{\nu}$	143000.0	58.4%	57.3%	11.9%
$e^-e^+ \to qq$	5897800.0	62.8%	13.2%	2.7%
$e^-e^+ \to qq\ell\nu$	11121800	28.3%	11.9%	0.3%
$e^-e^+ \to qq\ell\ell$	6639200.0	38.3%	2.9%	0.7%
$e^-e^+ \to qq\nu\nu$	2635000.0	71.4%	24.1%	5.3%
$e^-\gamma(BS) \to e^-qqqq$	2004388.1	23.3%	21.5%	0.8%
$e^+\gamma(BS) \to e^+qqqq$	2002334.1	23.4%	21.6%	0.8%
$e^-\gamma(EPA) \to e^-qqqq$	575600.0	12.0%	11.0%	0.5%
$e^+\gamma(EPA) \to e^+qqqq$	575600.0	12.0%	10.9%	0.4%
$e^-\gamma(BS) \to \nu qqqq$	414750.0	61.7%	59.5%	20.4%
$e^+\gamma(BS) \to \bar{\nu} qqqq$	414434.0	61.2%	59.1%	19.4%
$e^-\gamma(EPA) \to \nu qqqq$	108400.0	30.9%	29.9%	9.6%
$e^+\gamma(EPA) \to \bar{\nu} qqqq$	108400.0	30.7%	29.7%	9.1%
$e^-\gamma(BS) \to qqH\nu\nu$	92588.0	58.3%	56.2%	37.3%
$e^+\gamma(BS) \to qqH\nu\nu$	92430.0	58.1%	56.0%	37.1%
$e^-\gamma(EPA) \to qqH\nu\nu$	23400.0	30.1%	29.2%	19.4%
$e^+\gamma(EPA) \to qqH\nu\nu$	23400.0	29.7%	28.6%	18.8%
$\gamma(BS)\gamma(BS) \to qqqq$	18009413.9	54.2%	49.2%	1.9%
$\gamma(BS)\gamma(EPA) \to qqqq$	3824548.1	33.5%	30.2%	1.2%
$\gamma(EPA)\gamma(BS) \to qqqq$	3828498.1	33.7%	30.3%	1.2%
$\gamma(EPA)\gamma(EPA) \to qqqq$	805400.0	22.0%	19.8%	0.8%

Table 6.6: List of signal and background samples with the corresponding expected number at $\sqrt{s} = 3\,\text{TeV}$, assuming a luminosity of 2000fb^{-1}. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.

6.5.2 Sanity cuts

A set of very loose cuts, aiming to reduce the range of some discriminative variables to increase the effectiveness of MVA. (See section **??** on MVA) These cuts are very loose and physics motivated.

For $\sqrt{s} = 1.4 \, \text{TeV}$, invariant masses for H_{bb}, H_{WW^*}, W, and HH are smaller than 500, 800, 200, and 1400 GeV, respectively.

For $\sqrt{s} = 3 \, \text{TeV}$, invariant masses for H_{bb}, H_{WW^*}, W, and HH are smaller than 500, 800, 200, and 3000 GeV, respectively.

The selection efficiencies after sanity cuts and other pre-selection cuts stated above, are listed in table **??**.

6.5.3 Mutually exclusive cuts for $HH \to b\overline{b}W^+W^-$ and $HH \to b\overline{b}b\overline{b}$

Since the analysis for $e^-e^+ \to HH\nu\overline{\nu}$ channel is divided into two subchannels, $HH \to b\overline{b}W^+W^- \to b\overline{b}qqqq$ and $HH \to b\overline{b}b\overline{b}$, it is convenient to divided samples, both signal and background, into two mutually exclusive sets. This will make combining subchaneels much easier, as correlations between subchannels do not need to be considered.

The most distinctive difference between two subchannels, is that they have different number of jets, and different number of b-jets in the final state. So variables related to number of b-jets or a number of jets are suitable for separating two subchannels.

Shown in figure 6.9, two subchannels can be clearly separated in the two dimensional parameter space. The optimal rectangular cuts were selected by scanning the two parameters, and maximising

$$\varepsilon = P(\text{subchannel}_1|\text{selection}) \times P(\text{subchannel}_2|\neg\text{selection}) \tag{6.7}$$

where **selection** represents the mutually exclusive cuts, ¬**selection** indicates the phase space not covered by the **selection**.

Variables tested includes $\Sigma B_{4\text{jets}}$, $\sum_1^3 B_{4\text{jets}}$, y_{34}, y_{45}, y_{56}, y_{67} and other related variables. The best separation was summarised in table 6.7.
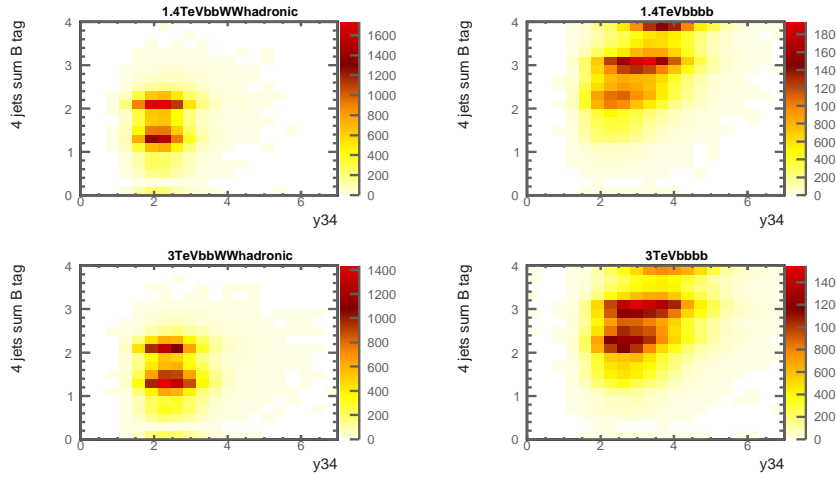
Figure 6.9: Sum of b tag against y_{34}, shown for signal samples

$\sqrt{s}$	selection	$HH \to b\bar{b}qqqq$ Selection Efficiency	$HH \to b\bar{b}b\bar{b}$ Selection Efficiency
1.4 TeV	$\Sigma B_{4jets} < 2.3$ and $y_{34} < 3.7$	86%	78%
3 TeV	$\Sigma B_{4jets} < 2.3$ and $y_{34} < 3.6$	89%	82%

Table 6.7: Mutually exclusive cuts, for full signal samples

The selection efficiencies after mutually exclusive cuts and other pre-selection cuts stated above, are listed in table **??**.

6.6 Discriminative Variables

A series of discriminative variables were calculated, and fed into MVA for signal selection.

The full list of variables can be found in table **??**. Same set of variables are used for $\sqrt{s} = 1.4$ TeV and $\sqrt{s} = 3$ TeV.

figure **??** hows the the variable XX which gives a good discrimination of signal against background.

The optimal set were chosen to give the best MVA performance, whilst no strong pair-wise correlation between any two variables, shown in figure **??**.

Channel / Efficiency	Sanity $\sqrt{s}$ = 1.4 TeV	Mutually exclusive $\sqrt{s}$ = 1.4 TeV	Sanity $\sqrt{s}$ = 3 TeV	Mutually exclusive $\sqrt{s}$ = 3 TeV
$e^-e^+ \to HH\nu\bar{\nu} \to$ $b\bar{b}W^+W^-\nu\bar{\nu}$, hadronic	66.4%	59.7%	69.5%	61.7%
$e^-e^+ \to HH\nu\bar{\nu} \to$ $b\bar{b}b\bar{b}\nu\bar{\nu}$	80.6%	15.4%	81.1%	18.8%
$e^-e^+ \to HH\nu\bar{\nu} \to$ other	24.7%	20.5%	25.1%	20.0%
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	42.0%	39.5%	40.3%	35.9%
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	43.4%	31.7%	39.2%	26.2%
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	64.2%	25.2%	64.2%	25.9%
$e^-e^+ \to qqqq$	4.6%	3.4%	2.5%	1.4%
$e^-e^+ \to qqqq\ell\ell$	3.3%	3.1%	0.7%	0.6%
$e^-e^+ \to qqqq\ell\nu$	11.4%.	9.8%	9.2%	7.2%
$e^-e^+ \to qqqq\nu\bar{\nu}$	18.8%	16.6%	11.8%	9.0%
$e^-e^+ \to qq$	2.0%	0.8%	2.5%	1.4%
$e^-e^+ \to qq\ell\nu$	1.6%	0.9%	0.3%	0.1%
$e^-e^+ \to qq\ell\ell$	0.1%	0.1%	0.7%	0.4%
$e^-e^+ \to qq\nu\nu$	5.8%	4.0%	5.3%	3.1%
$e^-\gamma(BS) \to e^- qqqq$	0.4%	0.3%	0.8%	0.7%
$e^+\gamma(BS) \to e^+ qqqq$	0.4%	0.4%	0.8%	0.7%
$e^-\gamma(EPA) \to e^- qqqq$	0.3%	0.2%	0.4%	0.4%
$e^+\gamma(EPA) \to e^+ qqqq$	0.3%	0.3%	0.4%	0.3%
$e^-\gamma(BS) \to \nu qqqq$	27.7%	25.3%	20.3%	16.8%
$e^+\gamma(BS) \to \bar{\nu} qqqq$	27.3%	24.9%	19.3%	15.9%
$e^-\gamma(EPA) \to \nu qqqq$	13.9%	12.6%	9.4%	7.8%
$e^+\gamma(EPA) \to \bar{\nu} qqqq$	13.7%	12.3%	8.9%	7.3%
$e^-\gamma(BS) \to qqH\nu\nu$	34.5%	30.2%	37.2%	30.2%
$e^+\gamma(BS) \to qqH\nu\nu$	34.2%	30.3%	37.1%	30.2%
$e^-\gamma(EPA) \to qqH\nu\nu$	18.2%	16.0%	19.0%	15.7%
$e^+\gamma(EPA) \to qqH\nu\nu$	18.7%	16.4%	18.4%	15.2%
$\gamma(BS)\gamma(BS) \to qqqq$	0.3%	0.3%	1.9%	1.7%
$\gamma(BS)\gamma(EPA) \to qqqq$	0.4%	0.3%	1.1%	1.0%
$\gamma(EPA)\gamma(BS) \to qqqq$	0.4%	0.3%	1.1%	1.0%
$\gamma(EPA)\gamma(EPA) \to qqqq$	0.4%	0.3%	0.7%	0.6%

Table 6.8: List of signal and background samples with the corresponding expected number at $\sqrt{s} = 1.4$ TeV and $\sqrt{s} = 3$ TeV, assuming a luminosity of 1500 and 2000fb^{-1}, respectively. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.

Variable	Description
$m_{H_{bb}}$	Invariant mass of H_{bb}
$m_{H_{WW^*}}$	Invariant mass of H_{WW^*}
m_W	Invariant mass of W
m_{HH}	Invariant mass of HH
E_{W^*}	Energy of W^*
E_{mis}	Missing energy, assuming collision at $\sqrt{s}$
$p_{TH_{bb}}$	Transverse momentum of H_{bb}
$p_{TH_{WW^*}}$	Transverse momentum of H_{WW^*}
p_{THH}	Transverse momentum of HH
η_{mis}	Pseudorapidity of missing momentum, assuming collision at $\sqrt{s}$
p_{THH}	Transverse momentum of HH
$-\ln(y_{23})$	minus ln of y_{23}. See section **??** for y parameter. See section **??** for the ln transformation
$-\ln(y_{34})$	minus ln of y_{34}.
$-\ln(y_{45})$	minus ln of y_{45}.
$-\ln(y_{56})$	minus ln of y_{56}.
$B_{1,H_{bb}}$	Highest b-jet tag value of two jets forming H_{bb}.
$B_{2,H_{bb}}$	Lowest b-jet tag value of two jets forming H_{bb}.
$B_{1,W}$	Highest b-jet tag value of two jets forming W.
B_{1,W^*}	Highest b-jet tag value of two jets forming W^* .
$C_{1,H_{bb}}$	Highest c-jet tag value of two jets forming H_{bb}.
$C_{1,W}$	Highest c-jet tag value of two jets forming W.
$|\mathbf{S}|$	Modulus of sphericity, $\mathbf{S}$. See section **??**.
$acol_{H_{bb}}$	Acolinearity of two jets forming H_{bb}.
$acol_W$	Acolinearity of two jets forming W.
$acol_{HH}$	Acolinearity of H_{bb} and H_{WW^*}.
$N_{H_{bb}}$	Number of PFOs forming H_{bb}.
$N_{H_{WW^*}}$	Number of PFOs forming H_{WW^*}.
N_W	Number of PFOs forming W.
N_{W^*}	Number of PFOs forming W^*.
$\cos\left(\theta^*_{H_{bb}}\right)$	Cosine of opening angles of two jets forming H_{bb}, in their rest frame.
$\cos\left(\theta^*_{H_{WW^*}}\right)$	Cosine of opening angles of W and W^*, forming H_{WW^*}, in their rest frame.
$\cos\left(\theta^*_W\right)$	Cosine of opening angles of two jets forming W, in their rest frame.

6.7 Multivariate analysis

Multivariate analysis was performed with TMVA package. The classifier that performs the best was found to be the boosted decision tree. See section **??** for details on boosted decision tree.

The parameters for boosted decision tree were optimised and checked for overtraining. The most important variables are the depth of the tree and the number of trees. Other parameters includes the minimum number of nodes in a leaf, the number of cuts of a variable, the learning rate, the sampling fraction, the yes/no or purity leaf, adaBoost or gradient boost.

The optimisation and overtraining test was done with $\sqrt{s} = 3\,\text{TeV}$ samples. $\sqrt{s} = 1.4\,\text{TeV}$ samples produce similar results.

Half of the samples were used for training, and the other half used for testing.

6.8 Signal selection results

6.9 Couplings extration

Colophon

This thesis was made in LaTeX 2$_\varepsilon$ using the "hepthesis" class [13].

Bibliography

[1] F. Gaede, Nucl. Instrum. Meth. **A559**, 177 (2006).

[2] J. S. Marshall, A. MÍźnnich, and M. A. Thomson, Nucl. Instrum. Meth. **A700**, 153 (2013), 1209.4039.

[3] W. Kilian, T. Ohl, and J. Reuter, European Physical Journal C **71** (2011).

[4] T. Sjostrand, (1995), hep-ph/9508391.

[5] S. Jadach, Z. Was, R. Decker, and J. H. Kuhn, Comput. Phys. Commun. **76**, 361 (1993).

[6] P. Mora de Freitas and H. Videau, p. 623 (2002).

[7] GEANT4, S. Agostinelli *et al.*, Nucl.Instrum.Meth. **A506**, 250 (2003).

[8] F. Gaede and J. Engels, EUDET Report (2007).

[9] J. S. Marshall and M. A. Thomson, Eur. Phys. J. **C75**, 439 (2015), 1506.05348.

[10] B. Xu, Improvement of photon reconstruction in PandoraPFA, in *Proceedings, International Workshop on Future Linear Colliders (LCWS15): Whistler, B.C., Canada, November 02-06, 2015*, 2016, 1603.00013.

[11] Particle Data Group, K. A. Olive *et al.*, Chin. Phys. **C38**, 090001 (2014).

[12] A. MÍźnnich, CERN Report No. LCD-Note-2010-009, 2010 (unpublished).

[13] A. Buckley, The hepthesis LaTeX class.

List of figures

List of tables