Summer Student Report
CERN

Optimisation of the VH to $b\bar{b}+X$ Selection

Fabian Wilk

mail@fabianwilk.de

Supervised by: William Hamish Bell

Abstract

This report presents results from two separated, yet related studies both using truth data: First a cut optimisation study and its results are presented. This study aims to provide a numerically optimised set of cuts for the current 8 TeV and the upcoming 14 TeV analysis of the $WH \to l\nu b\bar{b}$ channel.

In the second part an analysis of the fat-jet substructure of the $WH \to l\nu b\bar{b}$ channel is presented. This study is aimed at finding a discriminator for signal-like events against background events; it is expected that for large energies, a resolved analysis will fail and thus one has to rely on a different approach (such as substructure analysis) for a complete event selection.

Contents

1 Introduction

The observation of a HIGGS-like boson in 2012 was one of the most exciting discoveries in particle physics in the last sixty years. However, finding a particle is only one part of the scientific process: one has to measure the properties of this particle. So far, there has been no significant observation of the HIGGS-like boson in the fermion decay channel. Since this is a predicted process in the standard model (SM) and some model extensions predict fermiophobic HIGGS bosons, it is of great importance to measure the decay to fermions.

The expected SM HIGGS boson branching ratios are shown in Figure 1, where clearly the decay to $b\bar{b}$ is the most dominant contribution for the assumed HIGGS mass $m_H \approx 126\,\text{GeV}$.

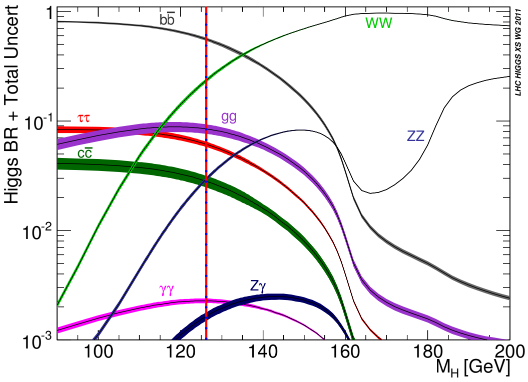


Fig. 1: Expected SM HIGGS boson branching ratio for low-mass regime [1].

The Decay Channel: An inclusive analysis of the bare $H \to b\bar{b}$ decay is not feasible, because of the enormous backgrounds, the high similarity of the signal and background signatures, and the intrinsic resolution problems when working with hadronic final states, just to name a few reasons. In order to reduce the background contributions, one does not study the bare $H \to b\bar{b}$ process but rather the associated production of a HIGGS boson with a vector boson, i.e. $VH \to b\bar{b} + X$, despite of the resulting cross section being more than a magnitude lower than the dominant gg fusion process [2].

By choosing a leptonic decay channel of the vector boson,[1] the resulting signal signature provides some separation power and the high p_T leptons produced in the process provide a means for efficient triggering and filtering of the process,

1.1 The Standard Set of Cuts

In this section the standard set of cuts for WH, used in this analysis and based on the most recent ATLAS VH study [2] are presented. In the resolved analysis, jets are clustered using the anti-k_t algorithm [3] with $R = 0.4$ as implemented by the FastJet [4] developers.

Listed below are the basic cuts on the object kinematics of the leptons, jets and the missing transverse energy.

Leptons		Jets					Missing transverse energy / W
Exactly one lepton with $p_T > 25\,\text{GeV}$		Exactly two b-tagged jets					Missing transverse energy from neutrino
Electrons	Muons	with	$p_T^{1\text{st}}$	$>$		$45\,\text{GeV}$	$E_T^{\text{Miss}} > 25\,\text{GeV}$
$\lvert\eta\rvert < 2.47$	$\lvert\eta\rvert < 2.5$		$p_T^{2\text{nd}}$	$>$		$25\,\text{GeV}$	
$\lvert\eta\rvert \notin [1.37, 1.52]$			$\lvert\eta\rvert$	$<$		2.5	
No additional leptons with $p_T > 15\,\text{GeV}$		No more than one additional jet					Transverse W-boson mass[2]
		with	$\lvert\eta\rvert < 2.5$, $p_T > 25\,\text{GeV}$	or		$2.5 < \lvert\eta\rvert < 4.5$, $p_T > 30\,\text{GeV}$	$m_T^W < 120\,\text{GeV}$

Due to the large energies at the LHC [5], it is expected that the decay products of the boson have a large boost factor. In order to exploit this feature, some cuts are applied in a binned approach, i.e. the cut values depend on the transverse momentum of the reconstructed vector boson p_T^V. One obvious example for such a cut is the $\Delta R(b,\bar{b})$ variable.[3] For the HIGGS boson signal, it is expected that for increasing energy (i.e. p_T^V) the decay products are more collimated and in turn one should apply a tighter cut on $\Delta R(b,\bar{b})$ in the highly boosted regime. The full list of binned cut variables and their cut values for each bin in p_T^V for the WH analysis are listed in Table 1.

1.2 Samples and Major Backgrounds

In this section the samples used in the analysis are presented and some additional information is given on how the respective background process emulates the signal.

Signal The signal sample was generated using Pythia8 [6] for a HIGGS mass of $m_H = 125\,\text{GeV}$.

Diboson These samples were generated using Herwig [7] for the diboson process and Jimmy [8] for the underlying event. The signature of the WZ diboson channel is extremely similar to the signal channel. Due to the

1 I.e. $X = \left\{l\nu, l\bar{l}, \nu\bar{\nu}\right\}$, however, this report concentrates on the WH mode, i.e. $X = l\nu$.

2 Defined by the ATLAS collaboration as $m_T^W := \sqrt{2p_T^l E_T^{\text{Miss}}[1-\cos(\phi^l - \phi^{\text{Miss}})]}$ with ϕ being the azimuthal angles of the charged lepton and missing transverse momentum [2].

3 Defined by the ATLAS collaboration as $\Delta R(b,\bar{b}) := \sqrt{\Delta\phi(b,\bar{b})^2 + \Delta\eta(b,\bar{b})^2}$ with $\Delta\phi$ being the difference in the azimuthal angle and $\Delta\eta$ the difference in η between the b and $\bar{b}$ system [2].

p_T^W [GeV]	0-90	90-120	120-160	160-200	> 200
$\Delta R(b,\bar{b})$	0.7-3.4	0.7-3.0	0.7-2.3	0.7-1.8	< 1.4
m_T^W [GeV]	≥ 40			≥ 0	
E_T^{Miss} [GeV]	> 25				> 50

Tab. 1: The set of topological cuts which is applied in different bins of the vector boson p_T.

fact that the Z-boson mass is relatively close to the HIGGS boson mass, there is very little possibility to reduce this background using kinematic or topological cuts. However, unlike the HIGGS boson, the Z-boson branching ratios to different fermionic final states differ by no more than one order of magnitude. Thus the number of WZ events that match the dual b-tag cut is much smaller than the total number of WZ events. The other diboson channels (WW and ZZ) are of lesser importance for the WH analysis.

$t\bar{t}$ The $t\bar{t}$ sample was simulated using `POWHEG` [9] and `Pythia6` [10]. A rather obvious reason for including this background is the fact that $t\bar{t}$ decays into two W-bosons and two b-quarks which have, in the mother top quark rest-frame, an energy of $\sim 65\,\mathrm{GeV}$, which is uncomfortably close to $m_H/2$ [11]. If one W-boson decays leptonically and the other is lost due to detector inefficiencies or acceptance, the event topology and kinematics are very similar to the signal. One approach for reducing the $t\bar{t}$ background is to apply cuts on the substructure of the $b\bar{b}$ fat-jet (cf. Section 3) because the signal $b\bar{b}$-system comes from a common decay vertex, whereas the $t\bar{t}$ produced $b\bar{b}$-system must be formed from two separate decay vertices.

W+Jets In this analysis only the heavy-flavour sub-sample was used, which was generated using `POWHEG` [9] and `Pythia8` [6]. The W+Jets and especially the heavy-flavour sub-sample is a major background, however, reduction can be achieved by requiring a dual b-tag and exploiting the fact that these b-jets are not necessarily produced in a common vertex for the background contribution. Typical approaches for background reduction are somewhat similar to the ones that reduce $t\bar{t}$ background.

1.3 Smearing of the Data Samples

In order to emulate the degradation of the input data due to detector resolution and inefficiencies and reconstruction effects one must apply a *smearing*, when using truth data. The effects of this smearing can be split in four categories:

Resolution In order to account for the limited detector resolution, all four-momentum components of each raw base particle type (jet, electron, muon, and missing transverse energy) are scaled. In the case of **electrons** and **jets** this scale factor is drawn randomly from a GAUSSian distribution with a mean of zero and a width that depends on the initial transverse momentum and η of the bare particle. The scaling factor for the $\boldsymbol{E_T^{\mathrm{Miss}}}$ is drawn randomly from a GAUSSian distribution with a mean of zero and a widths that depends on the sum of E_T of the neutrinos, the corresponding η and the E_T^{Miss} of the event; this smearing is applied after the neutrinos have been combined to form the E_T^{Miss}. For **muons**, the scaling factor is calculated in a deterministic way from the kinematic properties of the initial, bare muon.

Triggering It is required that a single electron or single muon trigger has fired. This trigger is simulated for each lepton individually and is composed of two steps: *1*) The trigger efficiency ε_T for electrons/muons based on their (smeared) kinematic properties is calculated; *2*) a number p is drawn randomly from a uniform distribution $[0, 1]$ and if $p < \varepsilon_T$ it is assumed that the trigger has fired.

Reconstruction The reconstruction/identification efficiency of electrons and muons is modelled; the method was already explained in the previous point (labelled *triggering*).

Tagging The smearing procedure for tagging has to account for two effects: *1*) **No-tags**, i.e. a *false negative* occurring when a true b-jet was not tagged, and *2*) **Mis-tags**, i.e. a *false positive* occurring when a c-/LQ-/τ-jet has accidentally received a tag.
Both are modelled by applying the procedure already explained in the *triggering* point. The former effect yields a reduction of b-tagged jets, whereas the latter introduces a non-b-jet contamination in the set of b-tagged jets.

The smearing factors and functions and the efficiencies have been obtained previously by the ATLAS collaboration using the full ATLAS detector simulation and an increased amount of interactions (as is expected for 14 TeV). This effort was done to provide an input to the European Strategy for Particle Physics.[4]

[4] See also the corresponding TWiki page [12]; the code used by this analysis uses the facilities provided by the `RootCore` package `TruthToRecoFunctions` that can be found in the `svn` repository [13].

2 Cut Optimisation

The main goal of the cut optimisation is to provide an updated set of cuts which yields in a more significant result to be used in the current 8 TeV analysis and, more importantly, by future 14 TeV analyses. It is crucial to include the smearing process explained in Section 1.3, because detector resolution and inefficiencies can severely influence the outcome of a selection.

2.1 Procedure

Following the object reconstruction and smearing as described in Section 1.3, the cut scan procedure is executed. This procedure performs a full event selection for the signal sample and all background contributions separately and calculates the resulting signal and background yields, N_S and N_B respectively. These yields are then plugged into Equation (1) which yields the significance ζ.

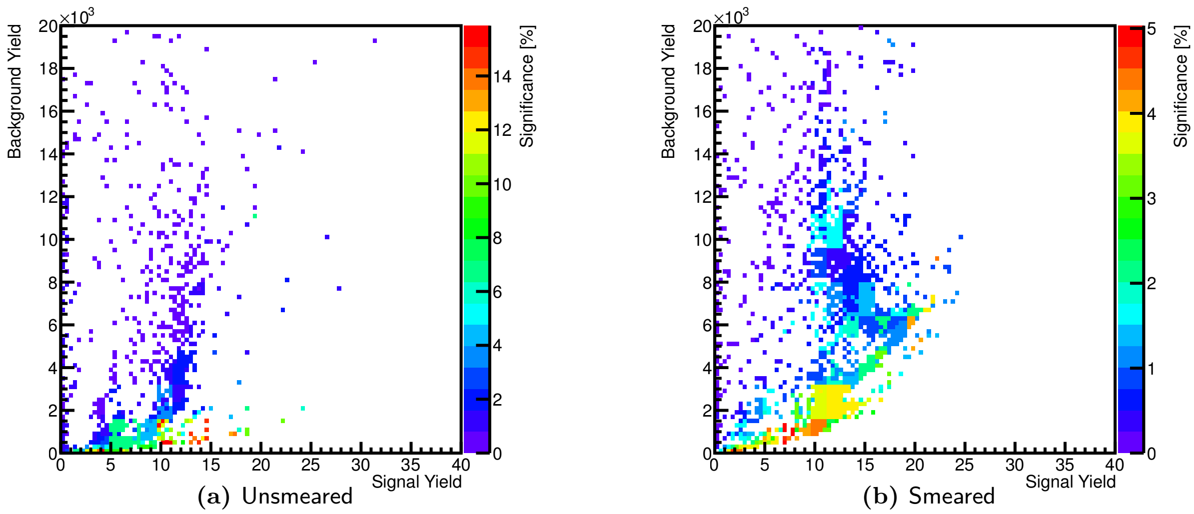
$$\zeta := \frac{N_S}{\sqrt{N_S + N_B + (0.3 \cdot N_B)^2}} \tag{1}$$

The significance ζ is a measure for the performance of the selected cut values. It should be stressed that possible systematic uncertainties are modelled by introducing a 30 % increase on the background yield obtained from the event selection. This is of special importance, because the modelling of the backgrounds is most likely not fully correct due to uncertainties and mis-modelling from MONTE-CARLO and the absence of some backgrounds (such as single-t).

A genetic algorithm was used to re-run the event selection procedure for a different set of cuts in each iteration. Based on the obtained value for the significance ζ, the cut values were modified within the space of possible cut values. As part of this study, the cuts on the p_T of the b-jets, the transverse W-mass m_T^W and the cuts on the $\Delta R(b,\bar{b})$ were varied using the genetic algorithm.

2.2 Results

Figure 2 shows the distributions of signal and backgrounds yields versus the resulting significance obtained by the cut optimisation procedure when run on the unsmeared (Fig. 2a) and smeared (Fig. 2b) input data. Obviously the significance is more than halved when going from unsmeared to smeared data. It should be noted that using smeared input data has an additional side effect: for smeared input data, the object reconstruction process is slightly changed; neutrinos and muons are removed from the list of objects used for jet clustering and therefore jets have lower energies.



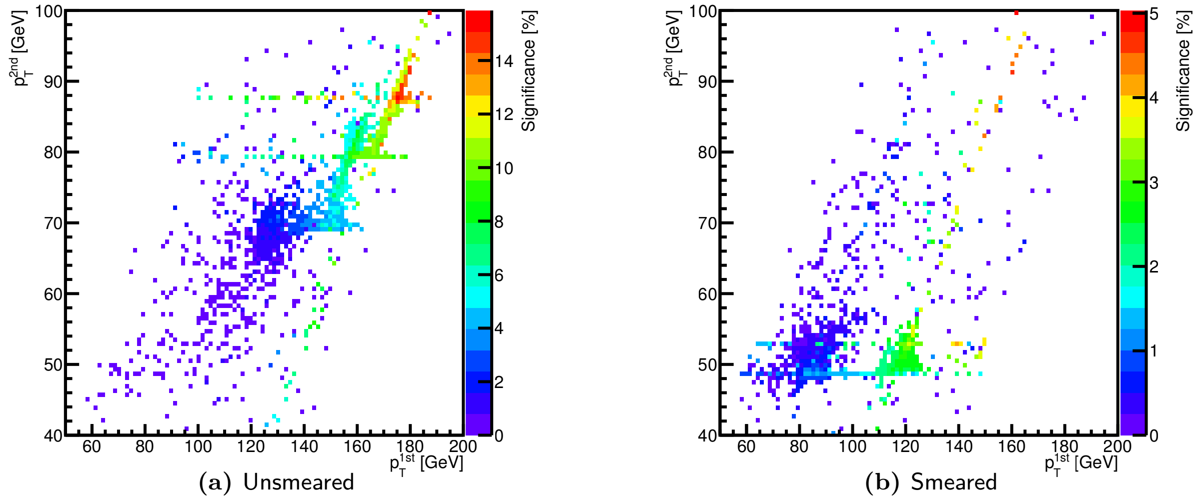
(a) Unsmeared **(b)** Smeared

Fig. 2: Signal Yields N_S vs. Backgrounds Yields N_B vs. Significance ζ for $\mathcal{L}_{\text{Int}} = 20.3\,\text{fb}^{-1}$ and $\sqrt{s} = 8\,\text{TeV}$.

Note that the axis labelling of the x- and y-axes differs by three orders of magnitude. The cut optimisation procedure shows that, in order to increase the significance one must use very tight cut values which result in a very small background yield but also in very little signal yield. For the unsmeared (smeared) input data, the maximal significance of $\zeta = 15.9\,(3.03)\,\%$ is obtained for $N_S = 3.99\,(1.99)$ and $N_B = 78.37\,(120.28)$. When comparing the values from the unsmeared and smeared input, it is obvious that the smearing degrades the separation power of the cuts which are used: although N_S decreases, N_B increases. Unfortunately, very low signal yields are somewhat problematic for two reasons:

- The simple definition of the significance ζ is problematic if $N_B \gg 1 > N_S$, because then $\zeta \sim 1/\sqrt{N_B + (0.3 \cdot N_B)^2}$ which is no longer sensitive to the signal component.

- It can be expected that there will be a relative increase in the background yield if one were to apply the same procedure on data samples because of background mis-modelling and, more importantly, the fact that some backgrounds (e.g. di-/multijet) are not modelled at all. Then, the signal yield might drop below an observable value, however, this will affect the lower bins of p_T^W more severely than the last bins.



(a) Unsmeared (b) Smeared

Fig. 3: p_T^{1st} vs. p_T^{2nd} vs. Significance ζ for $\mathcal{L}_{\text{Int}} = 20.3\,\text{fb}^{-1}$ and $\sqrt{s} = 8\,\text{TeV}$.

Figure 3 shows an example of a cut space map, in this case for the cuts on the b-jet p_T for the unsmeared (Fig. 3a) and smeared (Fig. 3b) input samples. Clearly a higher acceptance region for these cuts yields a larger significance. The hotspot of large significance is slightly lower for the smeared input, which matches the expectation.[5]

In addition to the aforementioned variables, cut value optimisation runs were performed varying the upper bound of the (unbinned) m_T^W cut and the lower and upper bound of the $\Delta R(b,\bar{b})$ cut (for each p_T^W bin separately).

$\boldsymbol{m_T^W}$ The optimisation prefers rather small values for this cut: the largest significance was found for a cut value around 45 GeV. However, this is most likely not a realistic result and caused by two aspects:
- The QCD (di-/multijet) background was not modelled, it is expected that an upper bound cut on m_T^W that is as small as 45 GeV would yield enormous background contamination.
- When using an upper bound cut $\sim$ 45 GeV, the influence of the low p_T^W bins is greatly reduced (because for the first three bins a lower bound cut $m_T^W > 40\,\text{GeV}$ is required). This feature was also observed in subsequent optimisation runs (for fixed cuts on m_T^W): the significance improves if the lower p_T^W bins (up to and including the third bin) are effectively disabled.

It should be noted that this cut optimisation was only performed using unsmeared input data.

$\boldsymbol{\Delta R(b,\bar{b})}$ Especially for the bins corresponding to large p_T^W, this cut is of great importance. The optimisation shows that low cut values increase the significance. For example, the best significance for the last bin corresponds to an acceptance range of $\Delta R(b,\bar{b}) \in [0.51, 1.01]$. This matches the expectation, because the high collimation of the two b-jets is a speciality of the boosted Higgs decay. Evaluation of a good lower bound cut on this variable is somewhat problematic, because the intrinsic scale that comes from the radius parameter R of the jet clustering introduces a bias (the two b-jets can not be closer than R because they must be clustered as two separate objects). This radius parameter was chosen to be $R = 0.4$, which is only a little smaller than the value extracted for the lower bound on $\Delta R(b,\bar{b})$ from the cut optimisation.

3 Substructure Analysis

At large energies, the b-jets coming from the highly boosted Higgs decay are collimated so much that a resolved analysis can no longer provide separation of the b-jets and thus fails at disentangling the final states; it is expected that this effect renders a resolved analysis impossible for $p_T^V > 200\,\text{GeV}$ [2]. This issue can be tackled by include the substructure of the jet objects in the event selection.

3.1 Basic Concepts

A substructure based analysis makes use of the fact that a reconstructed jet object is in fact a compound structure. It is composed of the physical objects which decay or shower hadronically (in the case of Monte-Carlo these are the simulated quarks, gluons, and hadrons; in the case of the experiment these are the pseudo-jets formed from

[5] As stated before, muons and neutrinos have been removed from the jet clustering and therefore the total jet energy is smaller.

firing calorimeter cells). As part of the reconstruction process, pseudo-jets are clustered to form a reconstructed jet which can be related to a physical object in the decay chain.

Instead of performing a fine-grained clustering of pseudo-jets (which is done by the resolved analysis in order to separate the two arms of a hadronic process) one can also combine pseudo-jet to much larger clusters. As a result, instead of obtaining jets that are the result of the hadronic decay in question, one obtains a *fat-jet* which contains the whole hadronically decaying particle. Consequentially one can analyse the composing objects of this fat-jet and devise a scheme for discriminating signal and background based on the *substructure* formed by the constituents.

In the substructure analysis the *Cambridge/Aachen* clustering algorithm [3] with radius parameter $R = 1.2$ was used; furthermore only the last p_T^W bin was considered.

3.2 The Mass Drop Tagger

The idea of using jet substructure variables as discriminator in HIGGS boson searches for associated production was first presented in [11], where a tagging scheme, called *mass drop tagging*, is proposed. This method is expected to provide reasonable separation power in the *boosted regime* i.e. when the HIGGS and the vector boson have large transverse momenta and are approximately back-to-back. The fundamental assumption is that a fast-moving HIGGS boson decaying to two b-quarks will appear as a single fat-jet that contains two highly collimated b-jets. In this analysis the implementation by the `FastJet` [4] developers was used. The basic concept of the mass drop tagger is to undo the jet clustering in an iterative manner and check in each step the mass drop, i.e. the ratio of the jet mass of the leading subjet and the jet mass of the compound jet.

A successful mass drop tag is usually followed by a filtering routine, which forms a fat-jet object at a small radius R_{Filt} such that it is composed of only the three hardest objects. This greatly reduces contamination from the underlying event (UE) while retaining the dominant hard QCD radiation [11]. These three objects are then assumed to be the two b-jets and a hard gluon.

3.3 Spatial Distribution of Constituents

A fundamental problem when working with the jet substructure is that one needs to find some parameterisation of the individual event which allows to combine the knowledge obtained separately for each event of a sample, in order to compare different types of samples. Naturally, each individual event will exhibit a different substructure: e.g. it is impossible to directly combine the spatial distributions of the fat-jet constituents of different events.

This analysis uses a parameterisation which transforms all fat-jet constituents to an event-independent coordinate system in the variables η and ϕ; this transformation is applied to each fat-jet constituent individually and is given by the following procedure:

1st Perform a translation $\mathcal{T}$ moving the position of the hardest fat-jet subjet to the coordinate origin, i.e. $x_1 = y_1 = 0$ for $p_{T,1} \geq p_{T,i} \quad \forall i \neq 1$.

2nd Introduce a scaling $\mathcal{S}$ which is constructed such that the distance between the hardest and the second-hardest fat-jet subjet is unity.

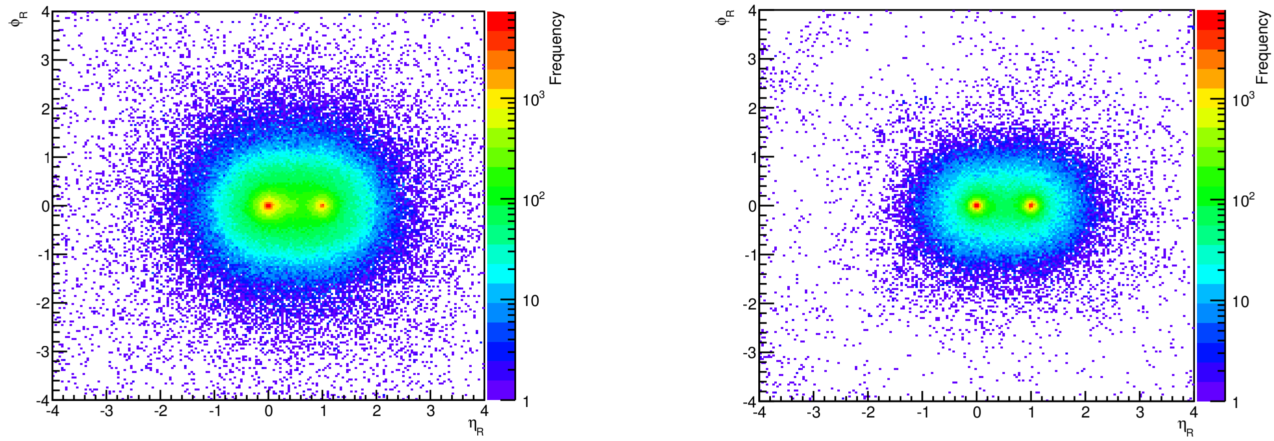
3rd Apply a rotation $\mathcal{R}$ whose rotation angle is chosen such that the second-hardest fat-jet subjet is rotated onto the positive arm of the x-axis, i.e. $x_1 = 0 < x_2$ for $p_{T,1} \geq p_{T,2} \geq p_{T,j} \quad \forall j > 2$.

After applying this set of transformations — i.e. $(\eta, \phi) \mapsto (\eta_R, \phi_R)$ — one can fill the coordinates for each fat-jet constituent for each event in one single two-dimensional histogram. The resulting 2-D distribution of fat-jet constituents is shown in Figure 4 for the lowest and highest bin in p_T^W as it was found for the $l\nu b\bar{b}$ signal sample.

Comparing the two figures, one can easily see that the fat-jets are much more collimated for larger p_T^W, which is in agreement with the expectation. The distribution in Fig. 4b is shaped such that the lengths of major and minor axes of the corresponding ellipse are much more asymmetric when compared to Fig. 4a. Looking very closely at the substructure in the direct vicinity of the hardest and second-hardest subjet one can see that for high p_T^W this region is in fact composed of two distinct circular regions rather than one cigar shaped segment.

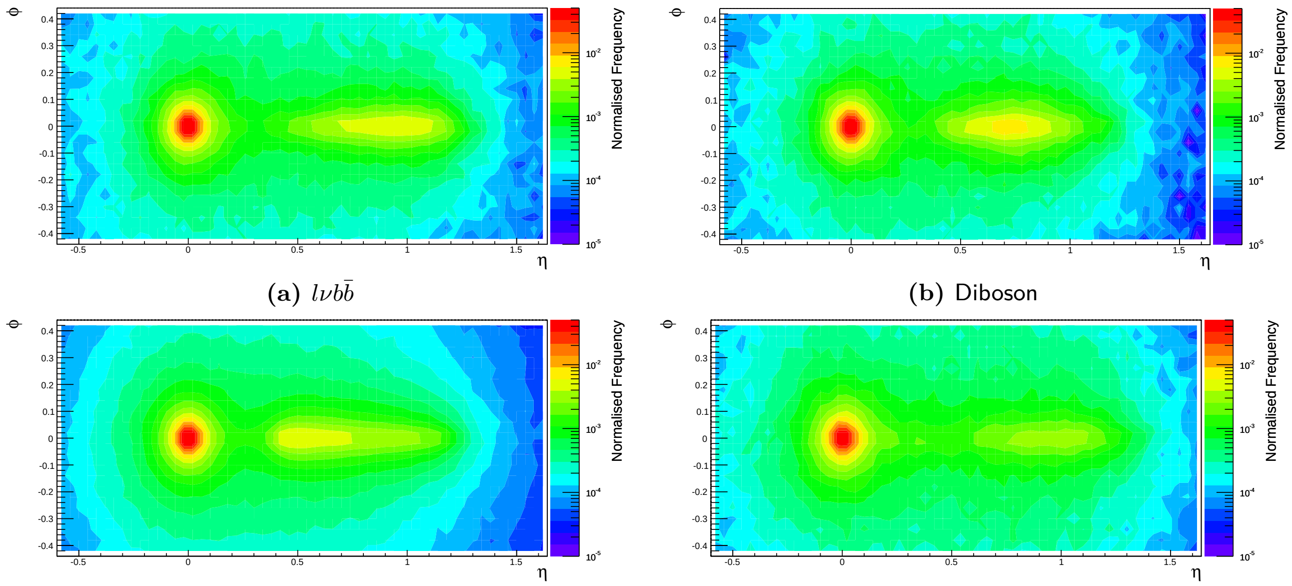
However, when using the above transformation rules, one loses a handle on the distances involved, due to the scaling step. By omitting the second step of the transformation rules one can obtain a distribution which retains the property of being event independent and gains the property that the distances are not re-scaled and thus correspond to the actual event distances. When omitting the scaling, one expects a large blob at $(0, 0)$ which corresponds to the hardest fat-jet subjet and a smeared out but still rather large contribution along the x-axis corresponding to the second-hardest fat-jet subjet.

Figure 5 shows a zoom-in to the region $(\eta, \phi) = [-0.6, 1.6] \times [-0.4, 0.4]$ where the coordinates have only been translated and rotated but not scaled; this is shown for the signal sample and the three most dominant backgrounds. Some observations are discussed below:



(a) Bin 0: $p_T^W \in [0, 90]\,\mathrm{GeV}$ **(b)** Bin 4: $p_T^W > 200\,\mathrm{GeV}$

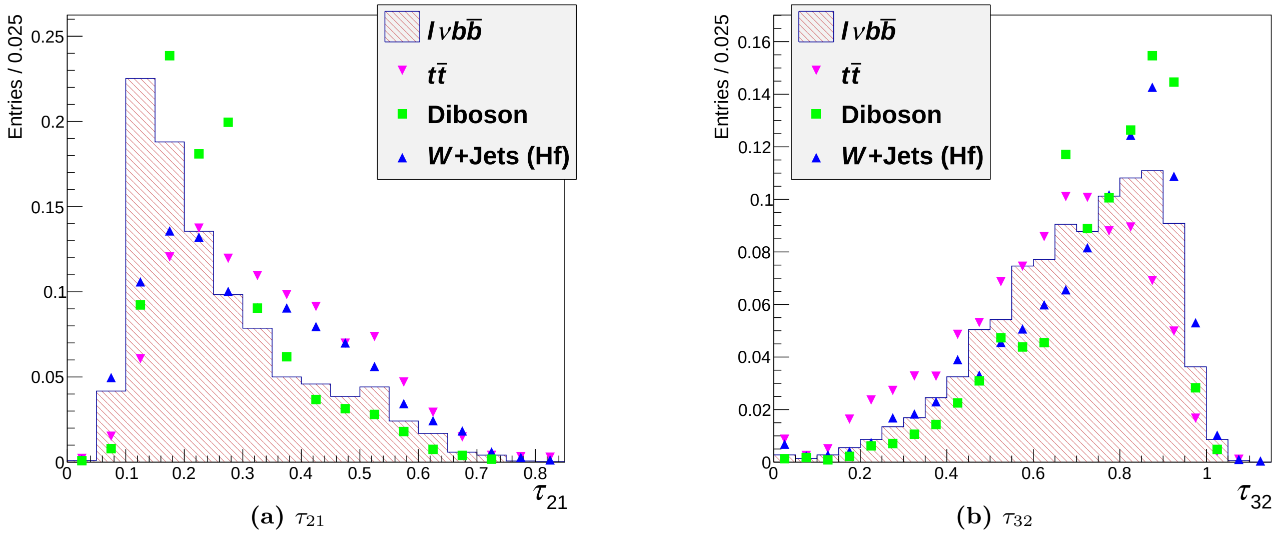
Fig. 4: Actual fat-jet substructure of events which pass mass drop tagging for the $l\nu b\bar{b}$ signal sample for the lowest and highest bin in p_T^W. Prior to filling to the histogram, each point was transformed using the transformation rules explained in Section 3.3. The left red blob corresponds to the hardest fat-jet subjet, the right red blob to the second-hardest fat-jet subjet. The analysis was performed for $\mathcal{L}_{\mathrm{Int}} = 20.3\,\mathrm{fb}^{-1}$ and $\sqrt{s} = 8\,\mathrm{TeV}$.



(a) $l\nu b\bar{b}$ **(b)** Diboson **(c)** W+Jets (Heavy-Flavour) **(d)** $t\bar{t}$

Fig. 5: Spatial distribution of fat-jet constituents in an unscaled, event independent coordinate system for the highest p_T^W bin for different samples. The histograms are normalised to unit area, the z-axes have been set to show the same interval for all samples. The analysis was performed for $\mathcal{L}_{\mathrm{Int}} = 20.3\,\mathrm{fb}^{-1}$ and $\sqrt{s} = 8\,\mathrm{TeV}$.

- All samples show a strong, highly significant peak at $(0, 0)$, which can be associated with the hardest fat-jet subjet.
- For signal, diboson and W+Jets one can clearly identify the structure resulting from the smeared second-hardest fat-jet subjet. In the case of the diboson sample, this structure is much more narrow and less smeared as compared to the signal sample. For W+Jets the structure is slightly closer to the coordinate origin (W+Jets: $\eta \in [0.4, 0.75]$, $l\nu b\bar{b}$: $\eta \in [0.65, 1.1]$).
- The $t\bar{t}$ background sample has a much more uniform distribution of the second-hardest fat-jet subjet; there is only a very small increase around $\eta \sim 1$. Generally speaking, it can be found that the whole distribution is much broader which must be attributed to a less collimated fat-jet.
- The right half of the distribution for W+Jets is not shaped like the radiation cone of a hard jet: The reduction of entries is much steeper than for the signal or diboson samples. Instead of a nearly circular contour shape *enclosing* the region of the second-hardest subjet, one can identify two, nearly straight edges. A possible explanation for this behaviour attributes this to the fact that the jet production is isotropic for W+Jets thus making it difficult to cluster two jets to a HIGGS-like fat-jet. Under this pretense, the second-hardest subjet might actually be hard radiation or fragmentation of the hardest subjet.



(a) τ_{21} (b) τ_{32}

Fig. 6: Distributions of subjettiness ratios; the distributions have been normalised to unit area.

3.4 Subjettiness Ratios

The Subjettiness τ_i is a measure for the likelihood that a certain (clustered) fat-jet has i real constituents (i.e. i subjet axes); it has been proposed that this variable might be of use for identifying boosted hadronically decaying objects [14, 15]. If τ_i is close to zero for a given fat-jet J, the fat-jet can be labelled i-jetty, i.e. it is likely that J has i subjets. Otherwise, if τ_i is larger, the opposite is true; typical values are $i = \{1, 2, 3\}$, for these $\tau_i \in [0, 1]$ [14, 15]. It was observed, that the actual i-subjettiness is not very stable with respect to sample fluctuations, however, the ratio $\tau_{ij} = \tau_i/\tau_j$ is much more stable.

Using the subjettiness as discriminator variable poses a special problem in the context of the mass drop tagging approach described in Section 3.2. After the tagging procedure has returned a positive tag, the fat-jet is filtered such that it contains only the three hardest objects as subjets. This is done using the `SelectorNHardest` routine implemented by the `FastJet` package, which filters the fat-jet to have *exactly* three subjets. However, the third subjet is only sensible when a hard gluon was radiated; the percentage of three-subjet event is expected to be much smaller than the observed 98 %.[6] By forcing the fat-jet to have three subjets, some bias is introduced and presumably some separation power lost.

One approach to solving this problem is to filter for the three hardest objects first, then check whether the softest object has a reasonable momentum (e.g. $p_T > 20\,\mathrm{GeV}$) and/or is well separated from the other objects and based on that either accept the three hardest objects or re-filter for two objects. Another approach involves substructure b-tagging; having obtained b-tag information on the three hardest objects allows to impose that the two hardest must have a b-tag whereas the third object should not have a b-tag.

Figure 6 shows the distributions of the subjettiness ratios τ_{21} and τ_{32} for the signal sample and the most dominant backgrounds; these results will be discussed below.

τ_{21} The samples are pairwise similar in shape: $t\bar{t}$ and W+Jets exhibit a similar shape and likewise do the signal and diboson sample. Specifically, one can see that low values of τ_{21} are preferred for the signal and diboson sample whereas such a precedence is not obvious for the W+Jets and $t\bar{t}$ samples. This is consistent with the expectation: only the diboson and signal signatures imply the existence of just two jet axes.

τ_{32} The diboson and W+Jets samples display a rather similar shape; the peak positions of the former two and the signal sample occur around the same value for τ_{32}; the peak heights of the $t\bar{t}$ and signal samples are similar. Although the distributions are not all alike, there seems to be no obvious separation power in the subjettiness ratio τ_{32}.

It should be stressed again that using the subjettiness ratios, after the filtering procedure was applied on the fat-jet, is somewhat prone to error because a bias might be introduced.

3.5 Fat-Jet Energy Ratio

[6] Although `FastJet` tries to filter to exactly three subjets, sometimes it fails which explains the deviation from 100 %.

Two new variables are introduced: the *In-Fat-Jet Energy* E_{IFJ} (*Close-To-Fat-Jet Energy* E_{CFJ}) is the sum of the p_T of all fat-jet constituents which fall in the green (red) hatched area in Figure 7; although these variables are sums of transverse momenta, they will be labelled as energies.

For the signal sample it is expected, that the majority of jet energy is in direct vicinity of the b-jets due to the fact, that the whole system is highly boosted. On the other hand, for a background sample such as the $t\bar{t}$ sample this is not necessarily true for obvious reasons. It is then expected, that the ratio E_{CFJ}/E_{IFJ} will tend to small values for the signal sample and large values for (some) background samples.

The resulting distributions for the last bin in p_T^W are shown in Figure 8; they were obtained using $R_{FJ} = 0.4$ which is slightly larger than the maximum filtering radius R_{Filt} that was used for reduction of UE and jet garbage (cf. Section 3.2). The shapes of the signal and diboson sample are relatively similar, both distributions show a very large maximum in the first bin and a very steep fall from there on. The $t\bar{t}$ and W+Jets samples on the other hand display a much smaller peak in the first bin (more severely for $t\bar{t}$) and fall much less steeply.

As was already the case with the subjettiness ratios (cf. Section 3.4), some discriminating power for rejecting $t\bar{t}$ and to a minor extent also W+Jets events by using substructure variables as part of an event selection can be identified. There has been no investigation concerning the impact of the shape of the E_{IFJ} and E_{CFJ} areas or the value of the parameter R_{FJ} on the separation power. One might study whether it helps if the shape is not *cigar*-like but rather a shape formed from two jointed circles.

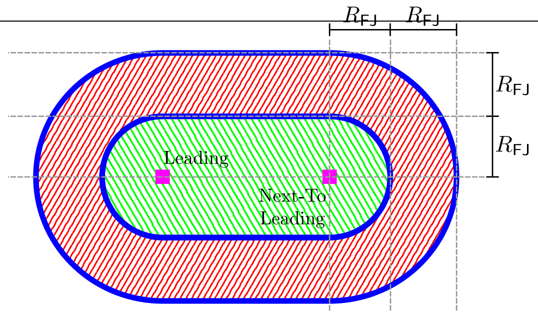


Fig. 7: Schematic drawing of the areas which are used for calculation of E_{IFJ} (hatched green) and E_{CFJ} (hatched red); R_{FJ} is a jet cone parameter (typical $R_{FJ} \sim 0.3$); magenta coloured squares denote position of leading and next-to leading fat-jet subjets.

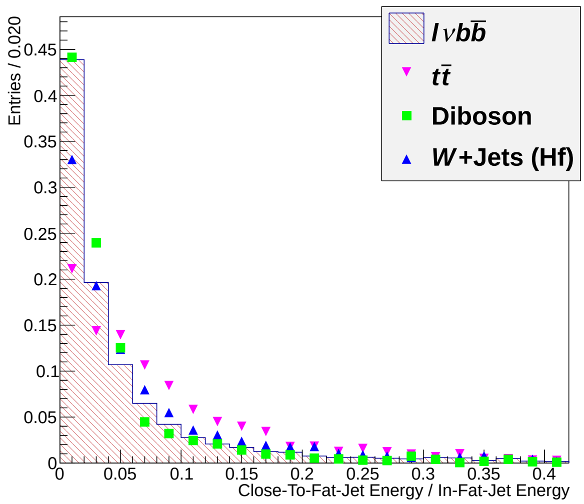


Fig. 8: Distribution of the fat-jet energy ratio for signal and background samples. Distributions have been normalised to unit area.

4 Conclusion

The cut optimisation has shown that tighter cuts on the minimal b-jet p_T and the $\Delta R(b,\bar{b})$ variable yield more significant results. However, tighter cuts also yield less signal events, therefore a subsequent study should investigate (using data which is closer to experimental data) whether the set of cuts obtained from the cut optimisation can actually be used on experimental data.

It was found that the bins corresponding to low p_T^W do not increase the significance of the analysis, therefore we recommend to drop these bins from the analysis. The most recent ATLAS analysis [2] includes them, however, mostly in order to get a handle on some backgrounds uncertainties and contributions. However, since the cut optimisation obtains a worse significance for the low p_T^W bins, we advise to use a different approach (for example by using control regions/sidebands) for handling the background. It might furthermore be worthwhile to include a cut on the $\Delta\phi(H,W)$ — which would ensure that the two arms of the event are roughly back-to-back — in the cut optimisation run; this cut is already implemented but currently disabled.

At the time of this writing, the 14 TeV data is not ready to be analysed. However, the tools for analysis and evaluation of the results are present and thus a future study can be built easily upon the tools and knowledge gained in this study. It might be expedient to include the W+Jets non-heavy-flavour sample and introduce some handle on QCD (di-/multijet) background. In addition, a subsequent study should further investigate the ZH channel and — instead of performing preliminary tests of the selection routines for this channel — also execute the cut optimisation procedure on the corresponding samples.

The substructure analysis shows some promising results, further investigation is necessary. The subjettines variables are somewhat biased due to the filtering procedure (cf. Section 3.4), therefore a subsequent analysis should investigate how this problem can be solved. Further development of the fat-jet energy ratio (cf. Section 3.5) can possibly yield a variable with nice separation power. It might be worthwhile to investigate how this variable performs for different values of R_{FJ} and for different shapes of the areas, e.g. by modelling the in-fat-jet area by joining two jet cones (forming a shape similar to a lemniscate) instead of filling the area between the two jet cones. Furthermore it might help to apply some additional cuts, such as requirements on the isolation of the b-jets or cuts on the jet mass, after the mass drop tag but before doing the subsequent substructure analysis.

A Bibliography

[1] A. Denner, S. Heinemeyer *et al.* *Standard Model Higgs-Boson Branching Ratios with Uncertainties*. *Eur.Phys.J.*, **C71**: p. 1753, **2011**. doi:10.1140/epjc/s10052-011-1753-8. ↑2

[2] *Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated W/ZH production with the ATLAS detector*, **July 2013**. ↑2, 5, 9

[3] Matteo Cacciari, Gavin P. Salam and Gregory Soyez. *The Anti-k_t jet clustering algorithm*. *JHEP*, **0804**: p. 063, **2008**. doi:10.1088/1126-6708/2008/04/063. 0802.1189. ↑2, 6

[4] Matteo Cacciari, Gavin P. Salam and Gregory Soyez. *FastJet User Manual*. *Eur.Phys.J.*, **C72**: p. 1896, **2012**. doi:10.1140/epjc/s10052-012-1896-2. 1111.6097. ↑2, 6

[5] Lyndon Evans, (ed.) and Philip Bryant, (ed.). *LHC Machine*. *JINST*, **3**: p. S08001, **2008**. doi: 10.1088/1748-0221/3/08/S08001. ↑2

[6] Torbjorn Sjostrand, Stephen Mrenna and Peter Z. Skands. *A Brief Introduction to PYTHIA 8.1*. *Comput.Phys.Commun.*, **178**: pp. 852–867, **2008**. doi:10.1016/j.cpc.2008.01.036. 0710.3820. ↑2, 3

[7] G. Corcella *et al.*. *HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)*. *JHEP*, **0101**: p. 010, **2001**. hep-ph/0011363. ↑2

[8] J. M. Butterworth, Jeffrey R. Forshaw and M. H. Seymour. *Multiparton interactions in photoproduction at HERA*. *Z.Phys.C*, **72**: pp. 637–646, **1996**. doi:10.1007/s002880050286. hep-ph/9601371. ↑2

[9] Carlo Oleari and Laura Reina. *$W^{\pm}b\bar{b}$ production in POWHEG*. *JHEP*, **1108**: p. 061, **2011**. doi: 10.1007/JHEP11(2011)040,10.1007/JHEP08(2011)061. 1105.4488. ↑3

[10] Torbjorn Sjostrand, Stephen Mrenna and Peter Skands. *PYTHIA 6.4 physics and manual*. *JHEP*, **05**: p. 026, **2006**. 0603175. ↑3

[11] Jonathan M. Butterworth, Adam R. Davison *et al.* *Jet substructure as a new Higgs search channel at the LHC*. *Phys.Rev.Lett.*, **100**: p. 242001, **2008**. doi:10.1103/PhysRevLett.100.242001. ↑3, 6

[12] ATLAS Collaboration. *ATLAS Input to the European Strategy for Particle Physics*, **September 2013**. URL https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/EuropeanStrategy. ↑3

[13] ATLAS Collaboration. *WebSVN: TruthToRecoFunctions*, **September 2013**. URL https://svnweb.cern.ch/cern/wsvn/atlasgroups/PhysicsAnalysis/EuropeanStrategy/TruthToRecoFunctions/. ↑3

[14] Jesse Thaler and Ken Van Tilburg. *Identifying Boosted Objects with N-subjettiness*. *JHEP*, **1103**: p. 015, **2011**. doi:10.1007/JHEP03(2011)015. ↑8

[15] Georges Aad *et al.* (ATLAS Collaboration). *Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment*. *JHEP*, **1205**: p. 128, **2012**. doi:10.1007/JHEP05(2012)128. 1203.4606. ↑8