B-tagging studies in $HH \rightarrow 4b$ analysis at $\sqrt{s} = 8$ TeV

Tijs Karskens - Summer student at CERN, CMS during summer 2013

Supervision by Andreas Hinzmann,
Maxime Gouzevitch & Maurizio Pierini

Short version intended for Summer Student Team

September 20, 2013

1 Introduction

Several Beyond the Standard Model (BSM) physics scenarios predict the existence of a new particle; either a spin 0 Radion, or an excited state of the Graviton (spin 2). This analysis looks for such a particle (refered to as X) in the $X \rightarrow HH \rightarrow 4b$ channel. The X (Radion, Graviton, or some other very massive particle) decays to two Higgs bosons, which in turn both decay to a b-quark and a b-antiquark. For these heavy X-particles (decaying nearly at rest in the labframe), the Higgs bosons will appear back-to-back, and very boosted. The jets resulting from the b-quark and b-antiquark will therefore merge, and the final state will therefore have two merged, *fat* jets (dijet state).

The goal of this analysis is to define the best searching channel for the modelled heavy resonance, and put an upper limit on its production cross section in the 1 TeV $< M_{res} <$ 3 TeV range. This paper is a (heavily) shortened version intended for the Summer Student team. A more extensive report will also be written, and uploaded within days.

2 Event selection

This analysis explores effective methods of finding a heavy resonance. To determine the effectiveness of each method, simulated Signal-events (which are simulated events with 100% branching ratio $X \rightarrow HH \rightarrow 4b$), are compared to regular Standard Model predctions in two different ways. First, graphs are compared to a simulated expected background absent signal (called QCD background). These events are generated using MADGRAPH and PYTHIA6. This suffices to understand our methods and algorithms, but is not precise enough for quantitative conclusions. For that reason, precise numbers will be constructed using a data-driven background estimation (see section 2.3).

To highlight the most interesting events for the analysis, several constraints are applied on all samples. Trigger (FAT750 OR HT650 OR HT750) is implemented, and is found to be 99% effective at a dijetmass > 890 GeV. In addition, each event must have at least two jets with $p_T > 30$ GeV and $|\eta| < 2.4$ for an accurate reconstruction, with a pruned mass close to the Higgs-mass. This mass window has been optimized, and is found to be between 110 and 135 GeV. Finally, a cut $\Delta\eta = |\eta_{jet1} - \eta_{jet2}| < 1.3$ is imposed on all events to reduce QCD background.

2.1 N-subjetiness

A possible discriminator is the N-subjetiness $\tau_{21} = \frac{\tau_2}{\tau_1}$, where

$$\tau_N = \sum_k P_T^k min(\Delta R_{1k}, \ldots, \Delta R_{Nk}) \tag{1}$$

So, assuming the jet is an N-pole, the minimum is taken over the N poles and summed over all jets. Therefore the smaller τ_N is, the closer the jet is to an N-pole; the smaller τ_{21}, the closer the jet is to a dipole (rather than monopole) structure.

2.2 B-tagging

B-tagging is a method used to identify jets originating from b-(anti)quarks, and is based on the lifetime of the decay products of the b-quark. Typically, hadrons containing b-quarks have a lifetime long enough to travel a measurable distance in the detector, yet still decay within the detector. So rather than all detected parts of the jet emerging from one point, this causes a 'secondary vertex' within the jet. The *Combined Secondary Vertex* algorithm combines secondary vertices and lifetime information to construct a probability discriminator (between 0 and 1), to distinguish b-quark jets from other jets. Two 'working points' are used; 'loose' (CSV discriminant > 0.244) and 'medium' (CSV discriminant > 0.679), with close to 10% and 1% mismatch respectively. Scale factors exist, but have not been applied, and result in an deviation less than 5% of true values.

2.3 Background estimation

The background is estimated using a data-driven ABCD method. A sideband is defined with a different jet mass window cut on the second jet (first jet remains in a $110 - 135$ GeV window). Then, the background for n b-tags is estimated using the spectrum of $n - 1$ b-tags.

Mass window	cut $n-1$	cut n
70 - 110 GeV	A	B
110 - 135 GeV	C	D

Table 1: Sideband method

Assuming $\frac{A}{C} \approx \frac{B}{D} \rightarrow D \approx \frac{A}{C} \cdot B$ (Table 1).

Two assumptions are made to be able to use this method. First, pruned jet mass and b-tags are assumed to be uncorrelated. As can be seen in figure 2.3, this assumption is false; an anti-correlation between pruned jet mass and b-tagging can be distinguished. Next, no contamination from background with real b-quarks ($\bar{b}$-quarks) like $t\bar{t}$ or VV is assumed (which is also untrue, yet unclear how much).

As a closure test, this is first done using Monte Carlo QCD.In general the estimations are not more than 1σ off, and is therefore considered enough for a rough first analysis. As QCD is only a first-order approximation to the real data, the actual background is then estimated using data events, using the same ABCD method.

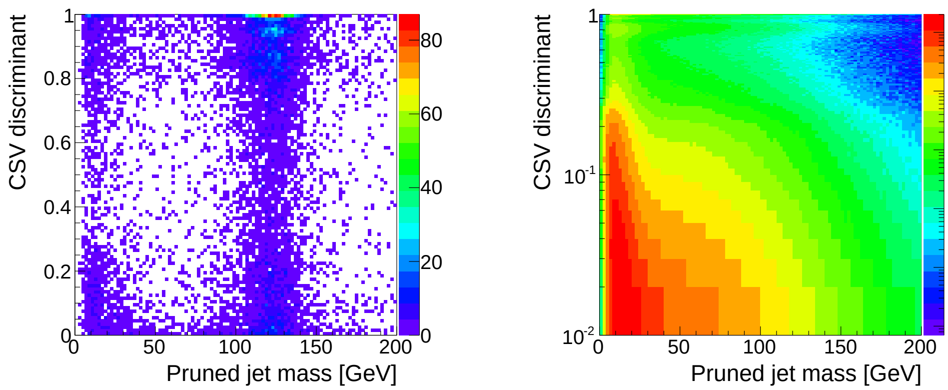

Figure 1: Pruned leading jet mass versus CSV disriminant of leading jet, leading subjet for signal (left) and QCD (right).

3 Results

3.1 Subjet b-tagging in boosted Higgs

Different b-tagging categories have been applied to the subjets. For each set the "Punzi significance", with $a = b = 2$, is used as estimator of its efficiency:

$$P = \frac{\epsilon_S}{1 + \sqrt{B}} \tag{2}$$

where ϵ_S is the signal efficiency, and B the number of background events remaining. Resulting numbers are shown in Table 2. It is clear that b-tags heavily reduce the background, while leaving most of the signal events. For loose b-tags, the significance increases the more b-tags are applied. For medium b-tags, 4 tags reduce the signal so much that it becomes less efficient.

3.2 N-subjetiness results

To test the effectiveness of N-subjetiness, a cut $\tau_{21} < 0.5$ on both jets is applied to all different b-tagging categories. Independent of b-tagging cut, the N-subjetiness reduces signal by a factor 1.3, and QCD roughly by a factor 3, and is therefore considered uncorrelated.

Cut	Signal	Background	Punzi significance
≥ 1 loose	2117 ± 46	7183 ± 53.6	0.001227
≥ 2 loose	2049 ± 45	1930 ± 27.6	0.002255
≥ 3 loose	1718 ± 41	298 ± 10.9	0.004582
exact 3 loose	879 ± 30	272 ± 10.4	0.002444
4 loose	839 ± 29	26 ± 2.9	0.006586
≥ 1 medium	2083 ± 46	3152 ± 35.5	0.001807
≥ 2 medium	1864 ± 43	509 ± 14.4	0.003873
≥ 3 medium	1186 ± 34	34 ± 3.5	0.008065
exact 3 medium	827 ± 29	31 ± 3.4	0.005823
4 medium	359 ± 19	3.0 ± 0.9	0.005762

Table 2: Punzi significance for different categories, using data-driven background estimation (1 TeV).

3.3 Background estimation improvements

So far only a provisory background estimation has been used. This has served well enough grasp the general effects of b-tagging and N-subjetiness in our boosted dijet topology, but is insufficient to put exact limits on the production cross section of a heavy resonance. For that reason, we have improved the initial background estimation (section 2.3). This is done by varying the parameters on *different* jets, rather than only using one jet in the sideband. The motivation behind this is that these parameters will be less correlated, but it is noteworthy that they are not completely uncorrelated either. Nevertheless, the estimation can work well enough if the correlation is sufficiently small. We use as signal region:

- Both jets $110 < m_{jet} < 135$
- Two different categories: 3 and 4 subjet b-tags

The following is used as method for background estimation, and results can be found in figure 2.

- Jet 1 randomly chosen, and has 1 or 2 subjet b-tags (for in total 3 b-tags or 4 b-tags respectively)
- Jet 1 has three mass regions: $110 < m_{jet1} < 135$ (signal region), $70 < m_{jet1} < 110$ (low mass-sideband) and $135 < m_{jet1} < 150$ (high mass-sideband)

- Jet 2 has $110 < m_{jet2} < 135$, and has three b-tagging regions: 0 subjet b-tags, 1 subjet b-tag and 2 subjet b-tags

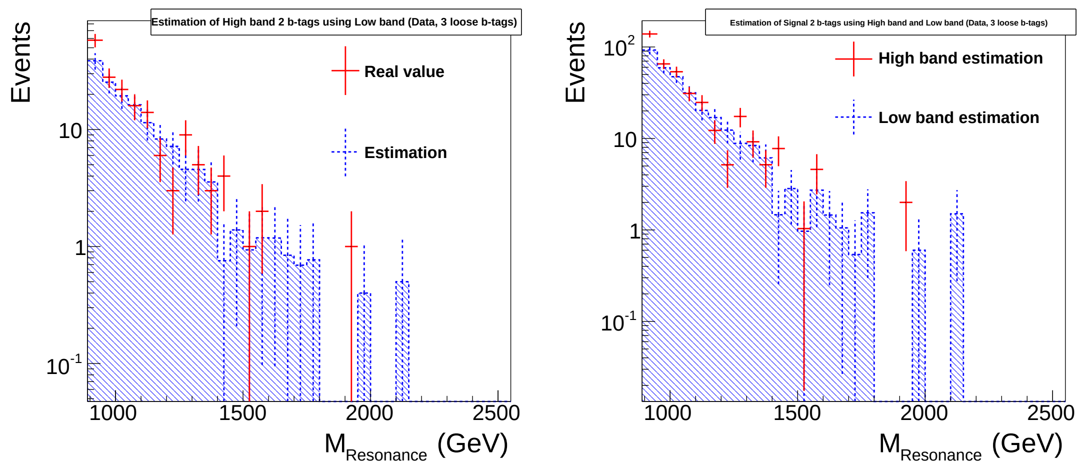

Figure 2: **Left:** Closure test, estimating high mass sideband using low mass sideband, using data. **Right:** Estimation of background, using high mass sideband and low mass sideband, using data.

4 Conclusions

Subjet b-tagging in a boosted dijet topology works very effectively as a background discriminator. Almost all QCD background events are removed, while leaving most of the signal intact. Categories with the highest efficiency are therefore 3 or 4 loose b-tags, and 3 or 4 medium b-tags. In addition, N-subjetiness applied on both jets ($\tau_{21} < 0.5$) appears to be an independent discriminator; regardless of applied b-tagging cuts, it reduces signal events roughly by a factor 1.3, while reducing QCD background events by a factor 3.

Because the simulated QCD background does not consider Electroweak bosons or top quark, this does not suffice as background. A background estimation using b-tagging and mass sidebands has been accomplished, and will be used in further analysis and limit setting on production cross section.