QCD VBF Jets+MET samples for Run 2

João Pela
on behalf of the Higgs to Invisible Analysis Group

Imperial College London

2015-07-29

Signal Topology: VBF Higgs to Invisible

- Two VBF jets with high mass but relatively low $p_\perp$ (~ 50 GeV)
- Reasonable amount of MET

Our ability to study QCD BG estimation methods is hampered by lack of MC statistics

QCD Private production during Run I

During Run I a set of QCD samples with VBF like jets and real MET was generated with:

- Real MET (sum os neutrino $p_\perp$)
- Two generator jets with high separation and mass

They allowed:

- Understand real MET in QCD
- Indirectly "determine the importance" of fake MET events (which we were not modelling)

This is not ideal because fake MET is important in our signal region

Proposal

We propose to generate samples of QCD with:

Filters

- Filter #1: VBF jet selection at GEN level
- Filter #2: MET applied after digitisation (implemented as a L1 MET cut)

By carefully tuning the cuts at each filter we could:

Advantages:

- Produce a QCD sample with high equivalent integrated luminosity (several fb^{-1})
- Generation cuts below trigger/offline analysis cuts
- Able to simulate both real QCD MET and fake/mismeasured MET
- If well understood possibility of using for MVA training

Assumptions and Questions

Assumptions

For this study it was assumed reasonable for full production:

- 5000 CPU for up to 2-3 months (60 to 90 days)
  - Is this a reasonable amount of computing power? If not which should be targeted?
- For an equivalent integrated luminosity of 10 $fb^{-}1$
  - If the computing requirements resources are too high a smaller equivalent luminosity sample could be made.
- The sample final size should be comparable to current produced samples

Expected QCD events

In order to determine which $p_\perp$ hats we would need to generate we calculate what are Phys14 QCD MC equivalent luminosities.

Equivalent luminosity per $p_\perp$ hat for Phys14 samples

pT Hat	X-Section (pb)	Events	Equi. Lumi (fb^{-1})
30-50	1.62×10^8	4844719	3.00×10^{-5}
50-80	2.21×10^7	4766292	2.16×10^{-4}
80-120	3.00×10^6	4795490	1.60×10^{-3}
120-170	4.93×10^5	4821602	9.78×10^{-3}
170-300	1.20×10^5	2794554	2.32×10^{-2}
300-470	7.48×10^3	2705941	3.62×10^{-1}
470-600	587.1	2925749	4.98
600-800	167	2857014	17.11
800-1000	28.25	499998	17.70
1000-1400	8.195	499986	61.01

Minimum $p_\perp$ to simulate

- For 10 fb^-1 we need to simulate up to bin 470-600
- For 30 fb^-1 we need to simulate up to bin 800-1000
- Above this $p_\perp$ hats we can use the QCD inclusive samples since they have enough equivalent luminosity.

Generator filter working points

The generator level filter will select events with at least a pair of generator jets (ak4) passing a set of cuts.

Generator Level Filter Statistics

						Expected times for 10 fb^{-1}		
	Filter Cuts					Events	CPU Times	
Selection	$p_\perp$	$\|\eta\|$	$\Delta\eta$	$\Delta\phi$	m_{jj}	Pass	HS06 CPU (s)	5k CPU (d)
Same as Run I without Generator MET	20	5.0	3.2	-	700	3.98×10^{10}	1.80×10^{13}	7031.31
Same as new HLT path	40	4.8	3.5	-	600	1.12×10^{10}	5.63×10^{12}	2195.87
Same as new HLT path + $\Delta\phi$	40	4.8	3.5	1.5	600	1.57×10^{9}	8.45×10^{11}	329.54
Working point A	50	4.8	3.5	1.5	1000	2.49×10^{8}	1.49×10^{11}	58.13
Working point B	50	4.8	3.0	2.0	1000	3.44×10^{8}	2.05×10^{11}	79.84

- All times are HS06 normalized (to reduce extrapolation error)
- Working Point A: Lowering $p_\perp$ even by 5 GeV increases processing time prohibitively
- Working Point A: Lowering $\Delta\eta$ cut makes almost no differences
- Working Point A: For each 100 GeV drop of m_{jj} will cost of at least 20 additional days.
- Increasing $min(\Delta\phi)$ to 2.0 and decreasing $\Delta\eta$ to 3.0 is acceptable (Working point B)

Post DIGI filter working points

For working point B ($p_\perp > 40$, $|eta| < 4.75$, $\Delta_\eta > 3.0$, $\Delta_\phi < 2.0$, $m_{jj} > 1000$)

Level 1 Cut

$p_\perp$ hat	Events 10fb-1	$L1_{ETM} \geq 40$	$L1_{ETM} \geq 60$	$L1_{ETM} \geq 70$
30-50	1.62×10^{12}	1.97×10^{6}	3.69×10^{5}	1.39×10^{5}
50-80	2.21×10^{11}	4.56×10^{7}	1.60×10^{7}	8.62×10^{6}
80-120	3.00×10^{10}	7.81×10^{7}	4.22×10^{7}	2.87×10^{7}
120-170	4.93×10^{9}	4.18×10^{7}	2.88×10^{7}	2.29×10^{7}
170-300	1.20×10^{9}	2.32×10^{7}	1.73×10^{7}	1.45×10^{7}
300-470	7.48×10^{7}	3.33×10^{6}	2.69×10^{6}	2.36×10^{6}
470-600	5.87×10^{6}	3.74×10^{5}	3.15×10^{5}	2.83×10^{5}
600-800	1.67×10^{6}	1.21×10^{5}	1.06×10^{5}	9.67×10^{4}
	Total	1.94×10^{8}	1.08×10^{8}	7.75×10^{7}

- Sample would have around 80-200M events for 10 fb^{-1} depending on L1 cut with this generator filter configuration
  - What would be an acceptable upper limit to the samples size?
- The preferred working point would be $L1_{ETM} \geq 60$ or lower if possible.

Technical requirements

If our proposal is accepted some technical requirements need to be fulfilled before central production can be started:

Requirements

- The generator dijet filter needs to be included in the CMSSW release.
  - Which version to target?
- The post-DIGI filter needs to be included in the CMSSW release.
  - To what package?
- cmsDriver needs to be updated to allow the inclusion of filters after DIGI.

Summary:

- A proposal of QCD samples with VBF jets and MET was presented.
- These samples would allow in depth study of QCD in the signal region of our analysis.
- To produce such samples some technical changes would be required.
- If produced this would be the first samples made with offline cuts in CMS.
- Would allow direct study of fake met due to miss measurement at high integrated luminosity for the first time.

Backup Slides

Run I QCD VBF Jets+MET samples

As a reminder here are the generator cuts used to generate the run I QCD samples.

MC Filter: Vectorial sum of neutrino E_T

- $\sum E_{\perp}(\vec{\nu}) > 40\ GeV$

MC Filter: Dijet Filter (AK5 GenJet no μ)

- Select jets with:
  - $p_{\perp} > 20\ GeV$
  - $|\eta| < 5.0$
- From selected jets at least one pair with:
  - $m_{jj} > 700\ GeV$
  - $\Delta\eta > 3.2$

Cross Sections

Cross Section for some QCD $p_\perp$ hats

	Cross Section [pb]		
$p_\perp$ hat [GeV]	8 TeV	13 TeV	Change
30-50	66285328	161500000.	+243.6%
50-80	8148778.0	22110000.	+271.3%
80-120	1033680.0	3000114.3	+290.2%
120-170	156293.3	493200.	+315.6%
170-300	34138.15	120300.	+352.4%
300-470	1759.549	7475.	+424.8%
470-600	113.8791	587.1	+515.5%
600-800	26.99	167.	+618.7%

As expected cross section for the this QCD $p_\perp$ hats increase significantly from 8 to 13 TeV.

Unfiltered production: Hardware

A single job for the whole simulation chain for 100 events was submitted to CERN lxbatch.

- GEN, SIM, DIGI, L1, DIGI2RAW, HLT:GRun
- RAW2DIGI, L1Reco, RECO

Hardware Step 1

$p_{\perp}$ hat	CPU Model	#CPU	RAM
30-50	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
50-80	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
80-120	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
120-170	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
170-300	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
300-470	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
470-600	Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz	32	65866004
600-800	Intel(R) Xeon(R) CPU L5640 @ 2.27GHz	24	30080044

Hardware Step 2

$p_{\perp}$ hat	CPU Model	#CPU	RAM
30-50	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
50-80	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
80-120	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
120-170	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
170-300	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
300-470	Intel(R) Xeon(R) CPU L5520 @ 2.27GHz	16	22003728
470-600	Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz	32	65866004
600-800	Intel(R) Xeon(R) CPU L5640 @ 2.27GHz	24	30080044

To note that at lxplus (which I will assume is representative of the grid resources) machines can be very different in terms of CPU and number of cores. Differences were observed in CPU time between machines executing exactly the same code of +50% (sometimes as high as +100%).

Unfiltered production: Computing times

Here are the statistics for the computing

Times Step 1

	Wall Time (s)					CPU Time (s)	
$p_\perp$ hat	Min Event	Max Event	Avg. Event	Total Loop	Total Job	Total Loop	Total Job
30-50	37.7242	319.636	151.967	15434.6	15450.4	10006.5	10010.4
50-80	79.1903	279.195	182.942	18493	18501.2	11542.1	11545.4
80-120	38.5559	383.707	158.552	16077.9	16120	12256	12259.4
120-170	68.5542	282.544	156.599	15815.5	15830.4	13143.3	13146.4
170-300	53.4359	420.627	179.195	18312.9	18401.6	15434	15436.8
300-470	106.444	532.75	262.03	26514.1	26533.9	20745.9	20753.4
470-600	67.0317	273.888	132.023	13309.9	13411.4	10321.1	10322.7
600-800	130.889	495.392	294.314	29769.9	29782.2	27509.9	27514.4

Time Step 2

	Wall Time (s)					CPU Time (s)	
$p_\perp$ hat	Min event	Max event	Avg event	Total loop	Total job	Total loop	Total job
30-50	1.84912	96.7409	8.95214	1077.07	1123.31	880.19	914.487
50-80	2.206	204.6	11.0235	1291.34	1333.55	1054.38	1082.67
80-120	6.79601	2324.04	80.6816	11755.2	13667.5	5881.53	7538.02
120-170	3.15147	107.821	9.90019	1159.06	1206.04	953.376	981.356
170-300	3.64276	105.52	12.397	1633.53	1674.87	1185.27	1211.87
300-470	4.382	114.562	13.9278	1582.28	1627.83	1356.5	1386.06
470-600	4.935	102.541	16.4371	1814.67	1859.81	1587.74	1616.06
600-800	6.15872	103.179	18.1286	1984.51	2029.99	1757.19	1784.77

Some conclusions

- It would be impossible to process every single event, it would take several millennia on a single CPU. We need some kind of gen filter.
- Average processing times fluctuate a lot over different machines. We should use normalized time.
- On average events take between 1 and 5 mintutes to pass step one and a few seconds to pass second step.

NOTE: I am including PU at average 35 interactions and 50ns bunch separation.

Unfiltered production: Normalized Time

I have decided to use normalized time (to some benchmark CPU power) for my calculation. Two units are normally used KSI2K and HS06 (default for the GRID).

Times Step 1

	Job	KSI2K		HS06	
$p_\perp$ hat	Time (s)	Factor	Time (s)	Factor	Time (s)
30-50	15796	1.06692	16853	4.16105	65728
50-80	18698	1.03829	19414	4.04931	75714
80-120	16350	1.26018	20604	4.9148	80357
120-170	15988	1.38135	22085	5.38742	86134
170-300	19442	1.33294	25915	5.19844	101068
300-470	27017	1.28704	34772	5.01958	135614
470-600	13516	3.44939	46622	13.4529	181829
600-800	30455	1.3539	41233	5.28018	160808

Time Step 2

	Job	KSI2K		HS06	
$p_\perp$ hat	Time (s)	Factor	Time (s)	Factor	Time (s)
30-50	1541	0.997404	1537	3.89163	5997
50-80	1564	1.34015	2096	5.22698	8175
80-120	14996	1.16618	17488	4.54821	68205
120-170	1418	1.33286	1890	5.20028	7374
170-300	2117	1.09542	2319	4.27208	9044
300-470	1813	1.42581	2585	5.56205	10084
470-600	2092	1.42447	2980	5.55641	11624
600-800	2239	1.46405	3278	5.71103	12787

Some conclusions

- Now normalized time scales up linearly with $p_\perp$ (as it should).
- We will be using HS06 for our calculations
- The average factors (to convert back) for the step 1 jobs is $factor_{KSI2K} = 1.5213$ and $factor_{HS06} = 5.9330$

Generator level working points

Filter statistics

					Expected times for 10 fb^{-1}					
Filter Cuts					CPU Times		HS06 Times			
$p_\perp$	$\|\eta\|$	$\Delta\eta$	$\Delta\phi$	m_{jj}	Time (s)	Time 5k (d)	Time (s)	Time 5k (d)	CERN 5k (d)	Notes
20	5.0	3.2	-	700	4.25×10^{12}	9848.75	1.80×10^{13}	4.17×10^{4}	7031.31	Same as Run I without Generator MET
40	4.8	3.5	-	600	1.28×10^{12}	2971.97	5.63×10^{12}	1.30×10^{4}	2195.87	Same as new HLT path
40	4.8	3.5	1.5	600	1.87×10^{11}	432.12	8.45×10^{11}	1.96×10^{3}	329.54	Same as new HLT path + $\Delta\phi$
50	4.8	3.5	1.5	1000	3.12×10^{10}	72.25	1.49×10^{11}	3.45×10^{2}	58.13	Working point A
45	4.8	3.5	1.5	1000	4.60×10^{10}	106.44	2.14×10^{11}	4.95×10^{2}	83.48	WP A: Lower 5 GeV Dijet $p_\perp$
40	4.8	3.5	1.5	1000	7.06×10^{10}	163.46	3.21×10^{11}	7.42×10^{2}	125.09	WP A: Lower 10 GeV Dijet $p_\perp$
50	4.8	3.25	1.5	1000	3.13×10^{10}	72.50	1.50×10^{11}	3.46×10^{2}	58.35	WP A: Lower 0.25 $\Delta\eta$
50	4.8	3.0	1.5	1000	3.14×10^{10}	72.63	1.50×10^{11}	3.47×10^{2}	58.48	WP A: Lower 0.50 $\Delta\eta$
50	4.8	3.5	1.5	900	3.93×10^{10}	90.90	1.87×10^{11}	4.33×10^{2}	73.00	WP A: Lower 100 m_{jj}
50	4.8	3.5	1.5	800	4.93×10^{10}	114.15	2.35×10^{11}	5.44×10^{2}	91.67	WP A: Lower 200 m_{jj}
50	4.8	3.5	1.5	700	6.28×10^{10}	145.40	2.99×10^{11}	6.92×10^{2}	116.62	WP A: Lower 300 m_{jj}
50	4.8	3.5	1.5	600	8.11×10^{10}	187.67	3.84×10^{11}	8.90×10^{2}	150.02	WP A: Lower 400 m_{jj}
50	4.8	3.5	1.5	500	1.04×10^{11}	241.42	4.92×10^{11}	1.14×10^{3}	192.10	WP A: Lower 500 m_{jj}
50	4.8	3.5	1.5	400	1.29×10^{11}	297.85	6.03×10^{11}	1.40×10^{3}	235.38	WP A: Lower 600 m_{jj}
50	4.8	3.5	1.75	1000	3.63×10^{10}	83.92	1.73×10^{11}	4.00×10^{2}	67.44	WP A: High 0.25 $min(\Delta\phi)$
50	4.8	3.5	2.0	1000	4.28×10^{10}	99.05	2.03×10^{11}	4.71×10^{2}	79.34	WP A: High 0.5 $min(\Delta\phi)$
50	4.8	3.5	2.25	1000	5.14×10^{10}	119.01	2.43×10^{11}	5.63×10^{2}	94.88	WP A: High 0.75 $min(\Delta\phi)$
50	4.8	3.5	2.5	1000	6.52×10^{10}	150.91	3.06×10^{11}	7.09×10^{2}	119.42	WP A: High 1.00 $min(\Delta\phi)$
50	4.8	3.5	2.75	1000	6.87×10^{10}	158.92	3.22×10^{11}	7.45×10^{2}	125.58	WP A: High 1.25 $min(\Delta\phi)$
50	4.8	3.5	-	1000	2.41×10^{11}	558.78	1.11×10^{12}	2.57×10^{3}	432.66	WP A: No $min(\Delta\phi)$ cut
50	4.8	3.25	2.0	1000	4.29×10^{10}	99.40	2.04×10^{11}	4.73×10^{2}	79.66	WP A: High 0.5 $min(\Delta\phi)$ and Lower 0.25 $\Delta\eta$
50	4.8	3.0	2.0	1000	4.30×10^{10}	99.60	2.05×10^{11}	4.74×10^{2}	79.84	Working point B

- Lowering $p_\perp$ even by 5 GeV increases processing time prohibitively
- Lowering $\Delta\eta$ cut makes almost no differences
- For each 100 GeV drop of m_{jj} will cost of at least 20 additional days.
- Increasing $min(\Delta\phi)$ to 2.0 and decreasing $\Delta\eta$ to 3.0 is acceptable (Working point B)