Study of the performance of the DeepTau "VSjet" discriminator in the 4tops analysis

Fabio Iemmi [1] Huiling Hua [1]
Hongbo Liao [1] Hideki Okawa [2] Yu Zhang [2]

[1]Institute of High Energy Physics (IHEP), Beijing

[2]Fudan University, Shanghai

May 12, 2021

Some nomenclature

- VSjet discriminator comes in 8 WPs, will use shorthands for them
  - VeryVeryTight → **VVT**
  - VeryTight → **VT**
  - Tight → **T**
  - Medium → **M**
  - Loose → **L**
  - VeryLoose → **VL**
  - VeryVeryLoose → **VVL**
  - VeryVeryVeryLoose → **VVVL**
- Signal sample: $t\bar{t}t\bar{t}$
- Background samples: QCD, $t\bar{t}$, $t\bar{t}$+X
- `HLT_PFJet450` is added to the signal triggers

Some nomenclature

- **For every WP**, I am looking at the following:
  - Number of signal events, **S**
  - Number of background events, **B**
  - Significance, $Z = \sqrt{2[(S+B)\log(1+S/B) - S]}$
  - Significance when a 20% unc. is associated with B, Z_{syst}
  - Reconstruction efficiency in $t\bar{t}t\bar{t}$, **E(sig)**: number of gen taus that are matched to a reco tau in signal sample
  - Reconstruction efficiency in background, **E(bkg)**: number of gen taus that are matched to a reco tau in background samples
  - Reconstruction purity in $t\bar{t}t\bar{t}$, **P(sig)**: number of reco taus that are matched to a gen tau in signal sample
  - Reconstruction purity in background, **P(bkg)**: number of reco taus that are matched to a gen tau in background samples

About Z_{syst}

- To compute Z_{syst}, **two options**:
  - `RooStats::NumberCountingUtils::BinomialExpZ (Double_t sExp, Double_t bExp, Double_t fractionalBUncertainty)` method
  - **Combine** datacard
- First method is fast but seemingly unstable under some conditions
- Second method requires some more work, but stabler
- Decided to go with Combine in the end
- Write the easiest datacard possible:
  - 1 bin (the WP), two processes (S, B)
  - 1 log-normal, systematic uncertainty affecting B
  - Set uncertainty to 20%
- Write 1 datacard for each WP and category, feed it to Combine

The datacard


```
imax 1  number of channels
jmax 1  number of backgrounds
kmax *  number of nuisance parameters

bin WP
observation -1

bin          WP          WP
process      S           B
process      0           1
rate         4.204       2633.654

bkg_unc lnN -            1.2
```

The Combine command

- To run Combine and have significances printed:

```
combine -M Significance datacard.txt -t -1 --expectSignal=1
```

- Meaning:
  - -M Significance: use "Signficance" method documented in Combine documentation
  - -t -1: fit an Asimov dataset
  - –expectSignal=1: generate Asimov dataset with $\mu = 1$

Before categories

	VVT	VT	T	M	L	VL	VVL	VVVL
S	14	19	25	32	43	63	89	91
B	89647	145460	232105	393839	709586	1618358	3063950	3115740
Z	0.0462	0.0490	0.0509	0.0509	0.0509	0.0499	0.0510	0.0517
E(sig)	0.2056	0.2674	0.3343	0.4065	0.4894	0.5843	0.6441	0.6528
E(bkg)	0.2556	0.3305	0.4097	0.4965	0.5930	0.7028	0.7670	0.7723
P(sig)	0.9147	0.8822	0.8400	0.7833	0.7028	0.5672	0.4444	0.4201
P(bkg)	0.3897	0.3110	0.2419	0.1729	0.1151	0.0602	0.0351	0.0341

1τ 0L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	4	6	8	10	13	20	27	23
B	2634	5619	9460	14165	25807	51725	92553	84908
Z	0.0819	0.0760	0.0772	0.0829	0.0826	0.0863	0.0892	0.0789
Z_{syst}	0.0074	0.0047	0.0037	0.0032	0.0023	0.0017	0.0012	0.0011
E(sig)	0.8851	0.8779	0.8727	0.8588	0.8341	0.7816	0.7113	0.7028
E(bkg)	0.9769	0.9728	0.9681	0.9599	0.9435	0.9088	0.8576	0.8524
P(sig)	0.9068	0.8696	0.8253	0.7606	0.6731	0.5278	0.3990	0.3922
P(bkg)	0.6652	0.4071	0.3030	0.2477	0.1634	0.0968	0.0586	0.0557

1τ 0L yields

	VVT	VT	T	M	L	VL	VVL	VVVL
$t\bar{t}t\bar{t}$	4	6	8	10	13	20	27	23
$t\bar{t}$	2190	3137	4476	6509	10081	18269	30013	26298
QCD	368	2378	4842	7461	15443	32989	61806	57951
$t\bar{t}$+X	75	103	142	194	283	467	734	660

1τ 1L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	3	5	6	8	10	15	21	19
B	839	1209	1728	2528	4028	7422	13164	12221
Z	0.1178	0.1325	0.1451	0.1561	0.1652	0.1793	0.1867	0.1710
Z_{syst}	0.019	0.018	0.016	0.014	0.012	0.0097	0.0075	0.0072
E(sig)	0.8786	0.8751	0.8613	0.8488	0.8222	0.7675	0.6947	0.6868
E(bkg)	0.9415	0.9302	0.9146	0.8903	0.8537	0.7723	0.6694	0.6688
P(sig)	0.9076	0.8720	0.8278	0.7681	0.6781	0.5339	0.4040	0.3984
P(bkg)	0.7461	0.6648	0.5779	0.4801	0.3611	0.2312	0.1405	0.1403

1τ 2L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	0.8	1.1	1.4	1.9	2.5	3.8	5.3	4.9
B	20.8	32.2	50.8	85.7	150.2	324.8	602.2	571.1
Z	0.1737	0.1898	0.1991	0.2007	0.2073	0.2120	0.2176	0.2045
Z_{syst}	0.13	0.13	0.12	0.10	0.084	0.061	0.047	0.046
E(sig)	0.8789	0.8693	0.8614	0.8447	0.8171	0.7563	0.6818	0.6782
E(bkg)	0.7366	0.6995	0.6351	0.5574	0.4328	0.2710	0.1727	0.1710
P(sig)	0.9029	0.8697	0.8178	0.7622	0.6679	0.5166	0.3893	0.3868
P(bkg)	0.2685	0.2160	0.1690	0.1256	0.0843	0.0449	0.0257	0.0255

1τ 3L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	0.0518	0.0726	0.0932	0.1233	0.1670	0.2720	0.4149	0.3837
B	0.3973	0.5690	0.7806	1.0087	1.5519	2.7000	4.3679	4.1507
Z	0.0805	0.0942	0.1035	0.1204	0.1318	0.1629	0.1955	0.1855
Z_{syst}	0.080	0.094	0.10	0.12	0.13	0.16	0.18	0.17
E(sig)	0.8977	0.8801	0.8700	0.8418	0.8115	0.7507	0.6533	0.6506
E(bkg)	0.8729	0.8589	0.8444	0.8270	0.8009	0.7783	0.6669	0.6673
P(sig)	0.9027	0.8379	0.7927	0.7224	0.6309	0.4638	0.3232	0.3269
P(bkg)	0.5399	0.5134	0.4570	0.4236	0.3378	0.2430	0.1604	0.1584

2τ 0L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	0.09	0.17	0.30	0.52	0.97	2.26	4.63	9.77
B	32.50	68.77	133.38	278.95	703.92	2934.12	10111.94	62923.10
Z	0.0166	0.0210	0.0264	0.0313	0.0364	0.0417	0.0461	0.0389
Z_{syst}	0.010	0.010	0.0098	0.0084	0.0063	0.0036	0.0021	0.00066
E(sig)	0.9695	0.9590	0.9634	0.9522	0.9422	0.9030	0.8599	0.7956
E(bkg)	0.9870	0.9881	0.9881	0.9862	0.9813	0.9647	0.9383	0.8970
P(sig)	0.9567	0.9503	0.9199	0.8772	0.8040	0.6715	0.5416	0.4296
P(bkg)	0.7867	0.7318	0.6863	0.6074	0.4672	0.2598	0.1415	0.0708

2τ 1L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	0.05	0.08	0.14	0.26	0.49	1.14	2.42	5.27
B	3.86	7.35	15.61	31.38	87.52	339.18	1030.43	3924.37
Z	0.0230	0.0307	0.0362	0.0463	0.0527	0.0620	0.0754	0.0842
Z_{syst}	0.021	0.026	0.027	0.032	0.023	0.015	0.011	0.0063
E(sig)	0.9729	0.9627	0.9575	0.9490	0.9347	0.9003	0.8427	0.7852
E(bkg)	0.8750	0.9012	0.9035	0.9066	0.8659	0.8179	0.7412	0.7030
P(sig)	0.9589	0.9422	0.9070	0.8565	0.7865	0.6564	0.5163	0.4122
P(bkg)	0.2959	0.3835	0.3642	0.3629	0.3130	0.2241	0.1476	0.1241

2τ 2L

	VVT	VT	T	M	L	VL	VVL	VVVL
S	0.004	0.008	0.016	0.028	0.053	0.133	0.341	0.827
B	0.105	0.164	0.265	0.452	0.944	3.515	10.357	48.405
Z	0.0133	0.0189	0.0310	0.0419	0.0540	0.0707	0.1052	0.1186
Z_{syst}	0.012	0.019	0.031	0.041	0.052	0.067	0.090	0.073
E(sig)	0.9710	0.9649	0.9537	0.9644	0.9062	0.8898	0.8052	0.7466
E(bkg)	0.7391	0.7370	0.7628	0.7499	0.7859	0.5851	0.3824	0.2246
P(sig)	0.9860	0.9296	0.8657	0.7911	0.7315	0.5904	0.4485	0.3586
P(bkg)	0.2013	0.1699	0.1969	0.1732	0.1694	0.1024	0.0528	0.0257

Some nomenclature

- **For every WP**, I am looking at the following:
  - **After Hideki's suggestion**: "gen-level efficiencies"
  - Define "gen-level categories" by making requirements on gen tau and gen leptons multiplicities
  - Ask for an equal number of reco taus
  - See how many of these reco taus are matched to the gen taus
  - $\varepsilon_{1m} = N^{\text{events}}_{\text{1-reco-tau-is-matched}} / N^{\text{events}}_{\text{total}}$
  - $\varepsilon_{2m} = N^{\text{events}}_{\text{2-reco-tau-are-matched}} / N^{\text{events}}_{\text{total}}$
  - $\varepsilon_{\text{compl}} = 1 - \varepsilon_{1m} - \varepsilon_{2m}$

1τ 0L

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.9687	0.9570	0.9403	0.9132	0.8707	0.7827	0.6794	0.6690
ε_{2m}	nan	nan	nan	nan	nan	nan	nan	nan
$\varepsilon_{\text{compl}}$	0.0313	0.0430	0.0597	0.0868	0.1293	0.2173	0.3206	0.3310

$1\tau\,1L$

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.9674	0.9559	0.9390	0.9182	0.8838	0.8123	0.7267	0.7171
ε_{2m}	nan	nan	nan	nan	nan	nan	nan	nan
$\varepsilon_{\text{compl}}$	0.0326	0.0441	0.0610	0.0818	0.1162	0.1877	0.2733	0.2829

1τ 2L

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.9622	0.9549	0.9395	0.9182	0.8932	0.8468	0.7794	0.7705
ε_{2m}	nan	nan	nan	nan	nan	nan	nan	nan
$\varepsilon_{\text{compl}}$	0.0378	0.0451	0.0605	0.0818	0.1068	0.1532	0.2206	0.2295

1τ 3L

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.9639	0.9484	0.9332	0.9198	0.9062	0.8765	0.8498	0.8532
ε_{2m}	nan	nan	nan	nan	nan	nan	nan	nan
$\varepsilon_{\text{compl}}$	0.0361	0.0516	0.0668	0.0802	0.0938	0.1235	0.1502	0.1468

2τ 0L

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.0117	0.0160	0.0263	0.0477	0.0642	0.1062	0.1424	0.1943
ε_{2m}	0.9883	0.9840	0.9726	0.9501	0.9318	0.8788	0.8204	0.7373
$\varepsilon_{\text{compl}}$	0.0000	0.0000	0.0011	0.0022	0.0040	0.0150	0.0371	0.0684

2τ 1L

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.0343	0.0369	0.0356	0.0610	0.0717	0.0944	0.1306	0.1870
ε_{2m}	0.9634	0.9589	0.9617	0.9350	0.9244	0.8995	0.8574	0.7740
$\varepsilon_{\text{compl}}$	0.0023	0.0043	0.0027	0.0040	0.0039	0.0061	0.0119	0.0390

2τ 2L

	VVT	VT	T	M	L	VL	VVL	VVVL
ε_{1m}	0.0668	0.0909	0.1163	0.0906	0.1247	0.1037	0.1241	0.1639
ε_{2m}	0.9332	0.9000	0.8779	0.9094	0.8692	0.8902	0.8586	0.8138
$\varepsilon_{\text{compl}}$	0.0000	0.0091	0.0058	0.0000	0.0061	0.0061	0.0173	0.0223

Conclusions

- Numbers looked odd at the beginning
  - We understand them better now, they seem OK
- **We used** to use the **Medium VSjet WP**
- In 5 out of 7 categories, the **VVL WP gives the best significance**
  - In the remaining two, VVVL WP gives the best significance
- **It seems we should go looser in VSjet**
  - **ZhangYu is seeing the same**