Data Driven Background Estimation in $H \to \tau_h \tau_h$

A short presentation by Gorm Galster,

inspired by the work done by:
Soshi Tsuno, Stefania Xella, Daniele Zanzi, Zinonas Zinonos

Outline of presentation

Motivation

The Needle and the Haystack

Estimating the Background

Some comments

Why data driven methods

- Avoid MC based systematics
- Avoid MC related statistic issue

Gorm Galster
gorm.galster@cern.ch

The SM (VBF) Higgs boson

- Second largest production cross section of SM H at LHC
- Event signature provides significant discrimination
  - Two tag jets in forward region
  - Higgs decay products in central region

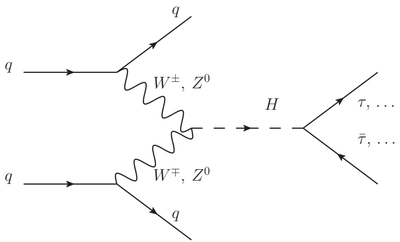


Figure 1: The VBF H production and decay.

The τ-lepton decay

- The hadronic decay mode is messy
- Decay is within beam pipe
- ν_τ carries away energy

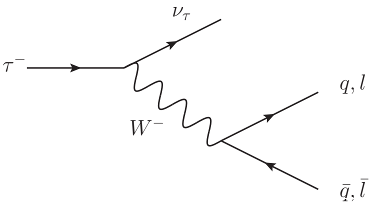


Figure 2: The τ-lepton decay. In $\sim 35\,\%$ of the times the τ decays leptonically, $l = e$, μ, in the remainder of cases the τ decay to a quark anti-quark pair, mainly u and d.

A neat visualization of the VBF signal

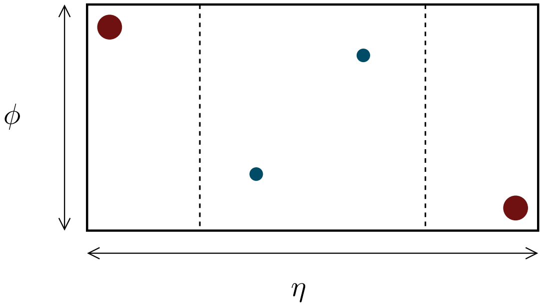


Figure 3: The event signature for VBF. In the forward region, in red, there are two tag jets, initiated by the production quarks. In the central region, in blue, two collimated jets from the tau decayτ.

The haystack

- Irreducible $Z \to \tau\,\tau$ (+ jets) background
- (Ir)reducible QCD background
- Other backgrounds
  - Z + jets
  - W + jets
  - $t\,\bar{t}$

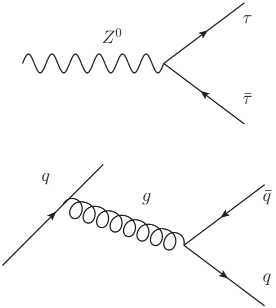


Figure 4: The dominant background. $Z \to \tau\tau$ (top) and QCD(bottom)

Gorm Galster
gorm.galster@cern.ch

An example cut flow

From control region to signal region:

- Background reduced by 40 %
- Signal reduced by 12 %
- Signal to background ratio improved by $\sim 46\,\%$

Cut $\int L\,dt = 4.16 fb^{-1}$	Data Events	$H(130) \rightarrow \tau_h \, \tau_h$ Events
All	459555648	2.41e+03±7
Cleaning	436472736	2.41e+03±7
Trigger	33675600	307±3
Lepton veto	30980472	296±2
2 BDT Loose	3903472	182±2
τe veto	698867	63.2±1
Good Run List	630152	63.2±1
τ-$p_T > 35, 20$ GeV	622366	62.5±1
$0 < x_1, x_2 < 1$	232993	32.6±0.8
$dR(\tau,\tau) < 2.6$	16971	11.3±0.5
$E_T^{\text{miss}} > 20$ GeV	6388	9.52±0.4
TTOS (control region)	933	7.42±0.4
$M(h,j) > 200$ GeV	652	6.9±0.4
$j_{lead} > 40$ GeV	561	6.55±0.3
TTOS (signal region)	561	6.55±0.3

Strategy for background estimation

If we for now assume control/signal region assume that...

- the shape of $Z \to \tau\tau$ background is unchanged
- the shape of QCD background is unchanged
- contamination in relevant distributions is negligible

... then

- $Z \to \tau\tau$ shape can be modelled in control region
- QCD shape can be modelled in control region (by requiring SS instead of OS)
- further constraints can be set...

Gorm Galster
gorm.galster@cern.ch

Strategy for background estimation

If we for now assume control/signal region assume that...

- the shape of $Z \to \tau\tau$ background is unchanged
- the shape of QCD background is unchanged
- contamination in relevant distributions is negligible

... then

- $Z \to \tau\tau$ shape can be modelled in control region
- QCD shape can be modelled in control region (by requiring SS instead of OS)
- further constraints can be set...

A QCD constraint term

Track multiplicity templates can be obtained

- for QCD: using SS control region
- for τ: from MC

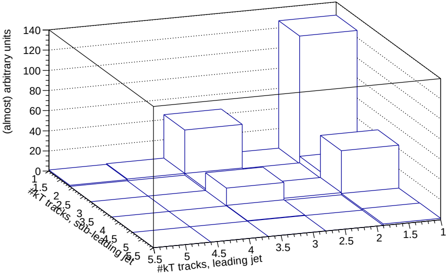


Figure 5: Number of tracks for leading and sub-leading for MC $Z \to \tau\tau$

Gorm Galster
gorm.galster@cern.ch

Implementation

The likelihood function for the system can be expressed as $\mathcal{L} = \mathcal{L}_{\text{control}} \cdot \mathcal{L}_{\text{signal}}$, where:

$$\begin{aligned} \mathcal{L}_{\text{control}} &= \mathcal{L}_{\text{shape}}(\tau \mid \vec{x},\, r_\tau) \times \mathcal{L}_{\text{shape}}(\text{QCD} \mid \vec{y},\, r_{\text{QCD}}) \\ &\times \mathcal{L}_{\text{tracks}}(\text{multiplicity} \mid r_{\text{exp}},\, r_\tau,\, r_{\text{QCD}}) \\ &\times G(N_{\text{obs}} \mid N_{\text{exp}} = r_{exp}\, N_{\text{obs}},\, \sqrt{N_{\text{exp}}}) \end{aligned} \tag{1}$$

$$\begin{aligned} \mathcal{L}_{\text{signal}} &= \mathcal{L}_{\text{shape}}(\tau \mid \vec{x}',\, r'_\tau) \times \mathcal{L}_{\text{shape}}(\text{QCD} \mid \vec{y}',\, r'_{\text{QCD}}) \\ &\times \mathcal{L}_{\text{tracks}}(\text{multiplicity} \mid r'_{\text{exp}}\, r'_\tau,\, r'_{\text{QCD}}) \\ &\times G(N'_{\text{obs}} \mid N'_{\text{exp}} = r'_{exp}\, N'_{\text{obs}},\, \sqrt{N'_{\text{exp}}}) \\ &\times G(\vec{x}',\, \vec{y}' \mid \vec{x},\, \vec{y}) \end{aligned} \tag{2}$$

Remarks and outlook

- Justification of assumptions made (some work already done by Daniele Zanzi)
- Cut flow
  - Definition of control region?
  - Definition of signal region?
- A relative measure
- Shape function, especially for τ, needs to be flexible
- Alternative/additional constraint on QCD? ABCD-method?
- Alternative formulation of likelihood function?

Gorm Galster
gorm.galster@cern.ch

Thank you.

Gorm Galster
gorm.galster@cern.ch

Back-up slides

oh, I should properly have said it any ways

Gorm Galster
gorm.galster@cern.ch

Z shape

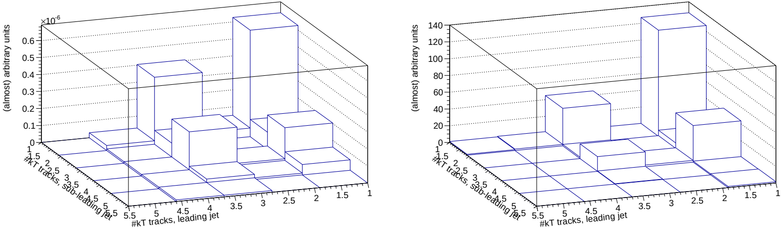
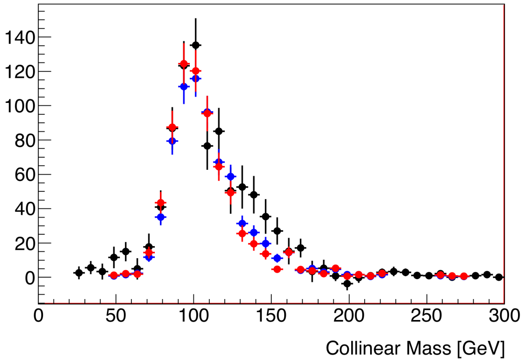


Figure 6: Number of tracks for leading and sub-leading for MC $Z \rightarrow \tau \, \tau$

Cut flow


140
120
100
80
60
40
20
0
0
50
100
150
200
250
300
Collinear Mass [GeV]

Ninjas

Figure 7: Ninjas. There five in this slide.

Gorm Galster
gorm.galster@cern.ch

No, there are no more slides. I am sorry.

Gorm Galster
gorm.galster@cern.ch