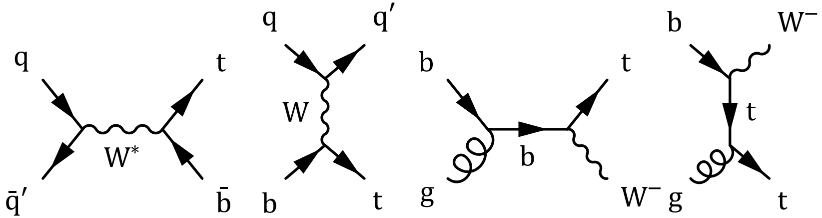
Standard Model Dilepton tZq Search at 13 TeV

Corin Hoad c.h@cern.ch

26th March 2018

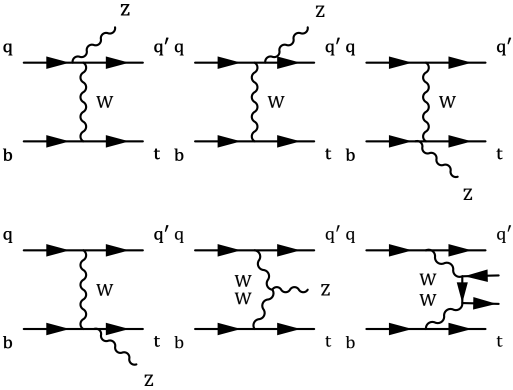
The top quark

- The top quark is the heaviest member of the standard model at $m_{\mathrm{t}} = 172\,\mathrm{GeV}/c^2$ and last quark to be discovered, in 1995.
- High $\mathcal{L}$ and $\sqrt{s}$ at the LHC allows probing of top quark interactions with unprecedented statistics.
- At the LHC, top quarks are predominantly created via. pair production as $\mathrm{t\bar{t}}$…
- …but single-top production via. the weak interaction is also possible:



Motivation

- Good test of SM predictions: sensitive to WWZ at a level on par with WZ production[1], and to tZ.
- Sensitive to BSM flavour changing neutral current couplings
- Background to other interesting, rare processes such as tHq



[1] doi:10.1103/PhysRevD.87.114006

Other Searches

Searches at 8 TeV during Run 1[2] and 13 TeV during Run 2[3] made for the trilepton final state.

- $t \rightarrow b + W,\ W \rightarrow l\nu,\ Z \rightarrow l\bar{l}$
- Cleaner topology than the dilepton channel, but with a smaller cross-section
- At 8 TeV $\sigma\,(tZq) = 10^{+8}_{-7}$ fb consistent with SM prediction of $8^{+0.7}_{-1.6}$ fb; observed with significance 2.4σ
- At 13 TeV $\sigma\,(tZq) = 123^{+33}_{-31}(\text{stat})^{+29}_{-23}(\text{syst})$ fb consistent with SM prediction of 94.2 ± 3.1 fb; observed with significance 2.7σ

[2]doi:10.1007/JHEP07(2017)003
[3]doi:10.1016/j.physletb.2018.02.025

Strategy

- Search for dilepton tZq events in 2016 pp collision CMS data using shape analysis
- Blinded analysis
- Verify background modelling with control regions rich in main backgrounds: Z/γ^{*} + jets (Drell-Yan) and $t\bar{t}$
- Impact of non-prompt electron and muon shapes gauged with data-driven estimates
- Multivariate analysis used to separate signal and background
- Fit performed to MVA response to measure cross-section and significance
- Full set of systematics included as nuisance paramters in fit

Backgrounds Considered

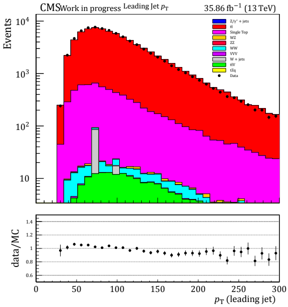
- **Single top prodution** t s-channel, t t-channel, tW, tHq
- **Top pair production** $t\bar{t}$, $t\bar{t}W$, $t\bar{t}Z$, $t\bar{t}H$
- **Boson+jets** Z/γ^{*} + jets, W + jets
- **Dibosonic** WW, WZ, ZZ, tWZ
- **Tribosonic** WWW, WWZ, WZZ, ZZZ
- Data-driven non-prompt lepton estimate

Event Reconstruction

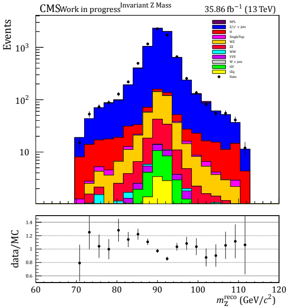
1. Exactly two, well identified and isolated leptons (ee or μμ) compatible with a Z considered.
    - Additional loose leptons vetoed.
2. At least four jets required: one from the top decay, two from the hadronic W decay and the recoil jet.
3. Up to two additional jets from gluon splitting are allowed.
4. At least one b tagged jet required
    - Top predominately decays to a b quark. One additional b jet from W decay/recoil quark is permitted.
5. W boson candidates reconstructed by choosing $\min \left| m_{\mathrm{j1}} m_{\mathrm{j2}} - m_{\mathrm{W}} \right|$.
    - Leading b jet assumed to originate from the t decay and so is excluded from consideration.

Control and Side Band Regions

- Two control regions (CRs) are established to compare the Monte Carlo (MC) simulation to the true data:
  - $t\bar{t}$ CR:
    - eμ dilepton final state with same invariant mass cut as signal Z mass cut (±20 GeV)
    - 1–2 b tagged jets
  - Z/γ^{*} + jets CR:
    - Currently identical to signal region, but requires 0 b tagged jets
    - New definition based on m_{W} and $\not{E}_{\mathrm{T}}$ being considered
- While blinded, data/MC comparisons are made in a side-band region.

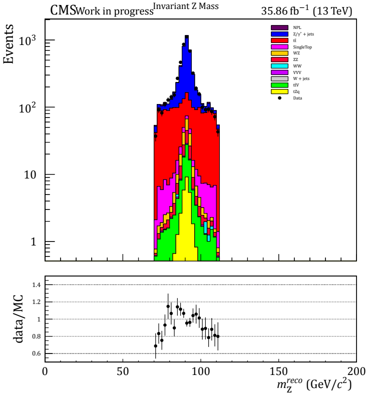


$t\bar{t}$ control region lead jet p_{T}



Z/γ^{*} + jets control region $m_{\mathrm{Z}}^{\mathrm{reco}}$

Data/MC Comparison for Signal Region in ee Channel



Reasonable agreement between MC and data is seen in the signal region for the ee channel.

Multivariate Analysis

- Multivariate analysis (MVA) techniques used after all other cuts and corrections to further distinguish signal from background.
- MVA techniques use multiple observables (features) to create a discriminator value that is greater for signal events.
- Different techniques considered:

Multi-layer perceptron Artificial neural network combining features arithmetically across a number of hidden layers to predict if an event is signal.

Boosted decision tree Series of decision trees formed by recursively splitting sample at a feature value which gives the best signal-background separation. Weights of incorrectly classified events increased for next tree in series.

Random forest Many decision trees trained on subset of available features. Discriminator taken as number of trees that consider an event to be signal.

Multivariate Analysis: Configuration

- Best-performing classifier found to be a BDT using the XGBoost[4] library
- XGBoost used to win the Kaggle Higgs Boson Machine Learning Challenge
- Configuration:
  - 75 trees
  - 0.075 learning rate
  - Stochastic boosting: only 67% of sample used in each tree
  - Max tree depth of 2
- Configuration optimised for ee channel; μμ channel will be configured independently.

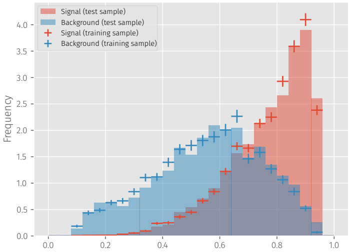
[4] doi:10.1145/2939672.2939785

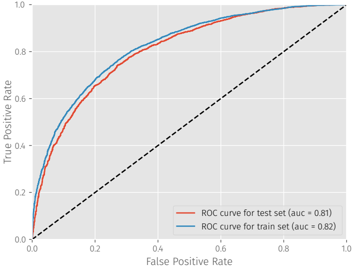
Multivariate Analysis: Features

Features used in BDT training in the ee channel in order of importance:

zMass	Reconstructed Z mass
topMass	Reconstructed mass of the top quark
jetMass	Total invariant mass of all jets
met	Missing transverse energy $\not{E}_{T}$
thirdJetPt	p_{T} of third most energetic
bTagDisc	b tag discriminant of the leading b jet
totHtoverPt	Ratio of total H_{T} to total p_{T}
jjDelR	ΔR of the two leading jets
leadJetPt	p_{T} of leading jet
leadJetEta	η of leading jet

MVA Response





Kolmogorov–Smirnov test signal (background): 0.89 (0.16)

Good separation between signal ad background and very little overtraining.

BDT Discriminant Binning

- To calculate significance and cross-section, BDT discriminant must be binned.
- Naïvely want to choose equidistant bin edges, it's easiest!
  - But how many?
  - Too many and statistics in individual bins is bad, too few and resolution of distribution is lost.
- Alternative: find the median of the data, and use it as a bin edge. Keep splitting the resulting bins at the median until a minimum number of signal or background events or a maximum percentage error is reached.
  - Min signal events: 0
  - Min background events: 1
  - Max bin error: 30%

Significance and Signal Strength

- Significances calculated using asymptotic approximation
- Signal strength calculated with maximum likelihood fit
- Systematic uncertainties incorporated as nuisance parameters
- Before unblinding, Asimov dataset used to calculate expected significance and signal strength

Summary

- Search for tZq events in 2016 pp collision events at CMS with $\sqrt{s} = 13\,\text{TeV}$.
- First analysis at CMS looking for tZq in the dilepton decay channel
- Control regions established to validate modelling
    - Good agreement between data and MC in control regions
- Thorough exploration of MVA techniques to achieve best separation
    - Current best-performing classifier is BDT using XGBoost library

Backup

Full Event Selection

- Signal electrons:
  - $p_T > 35$ GeV (leading)
  - $p_T > 15$ GeV (subleading)
  - $|\eta| < 2.5$
  - Identified as electron
- Signal muons:
  - $p_T > 26$ GeV (leading)
  - $p_T > 20$ GeV (subleading)
  - $\text{Iso}_{\text{rel,PF}}\,(\Delta\beta, R_{\text{Cone}} = 0.3) < 0.15$
  - $|\eta| < 2.4$
  - Identified as muon
- Jets:
  - $p_T > 40$ GeV
  - Lepton-jet separation $\Delta R < 0.4$
  - $|\eta| < 4.7$
  - Identified as jet
- Veto electrons:
  - $p_T > 35$ GeV (leading)
  - $p_T > 15$ GeV (subleading)
  - $|\eta| < 2.5$
  - Identified as electron
- Veto muons:
  - $p_T > 26$ GeV (leading)
  - $p_T > 20$ GeV (subleading)
  - $\text{Iso}_{\text{rel,PF}}\,(\Delta\beta, R_{\text{Cone}} = 0.3) < 0.25$
  - $|\eta| < 2.4$
  - Identified as muon
- b jets:
  - Identified as b jet
  - $|\eta| < 2.4$

- Additional requirements:
  - Exactly two tight leptons and no additional loose leptons
  - 4–6 selected jets
  - 1–2 b tagged jets
  - Z candidate leptons within $m_{ll} = m_Z \pm 20$ GeV
  - W candidate jet pair within $m_{jj} = m_W \pm 20$ GeV

Simulation Corrections

Various corrections to the simulation data are applied to match the data:

- Lepton ID, Isolation (muons) and reconstruction data/MC Scale Factors
- Electron energy and resolution Regression and Scaling and Smearing corrections
- Jet Energy Resolution data/MC Scale Factors
- b tagging scale factors
- Pileup modelling
- Trigger scale factors
- Rochester corrections

Background Estimation: Non-prompt leptons

- Background where at least one jet is incorrectly reconstructed as a lepton (predominately electrons) or lepton from a heavy quark decay (predominantly muons) are estimated with data.
- Use similar methodology to previous top quark pair production measurements[5] and same-sign SUSY searches[6]:
  - Most same-sign leptons are non-prompt or misidentified
  - Backgrounds are independent of the charge of the lepton pairs
  - …so we expect opposite-sign (OS) sample will have similar contribution to same-sign (SS)

Data-driven estimate is derived using:

$$N_{\text{data}}^{\text{OS fakes}} = \left(N_{\text{data}}^{\text{SS}} - N_{\text{expected real}}^{\text{SS}}\right) \frac{N_{\text{MC}}^{\text{OS fakes}}}{N_{\text{MC}}^{\text{SS fakes}}}$$

---

[5] doi:10.1140/epjc/s10052-017-4718-8

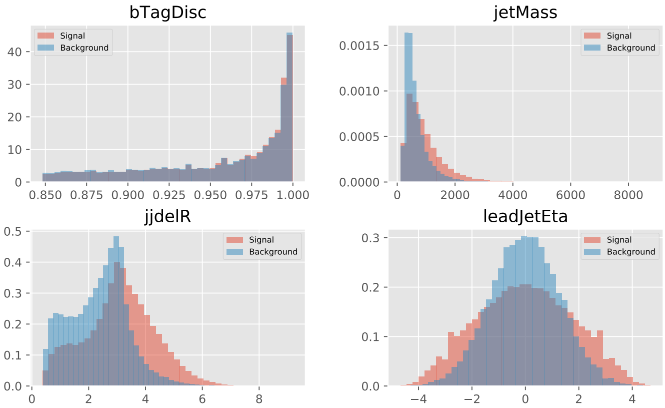
[6] doi:10.1140/epjc/s10052-016-4261-z

Side-band definition

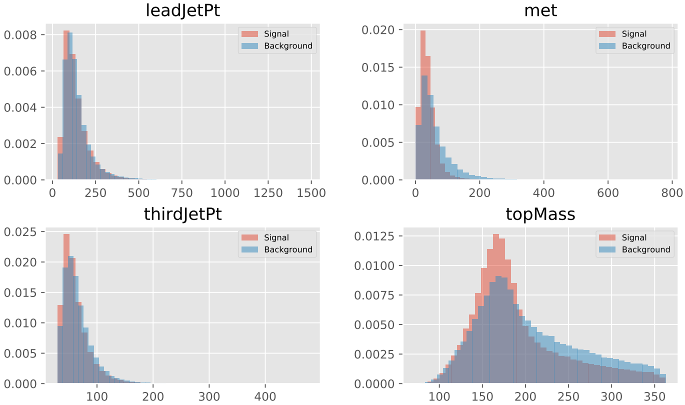
Signal region is defined as $\chi^2 < 40$ and side-band as $40 \leq \chi^2 < 150$, where

$$\chi^2 = \left(\frac{m_{\mathrm{t}}^{\mathrm{reco}} - m_{\mathrm{t}}}{\sigma_{\mathrm{t}}}\right)^2 + \left(\frac{m_{\mathrm{W}}^{\mathrm{reco}} - m_{\mathrm{W}}}{\sigma_{\mathrm{W}}}\right)^2.$$

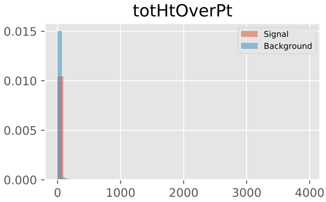
MVA Features: ee

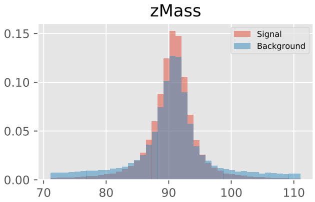


MVA Features: ee

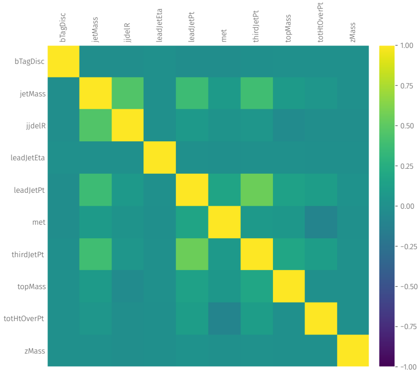


MVA Features: ee

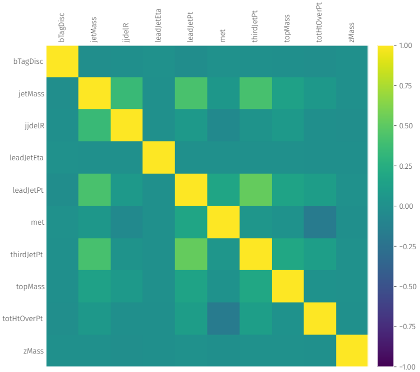




MVA Features: ee



Signal



Background

Little correlation between any features used in BDT training in the ee channel.

Systematic Uncertainties

Luminosity Rate uncertainty of $\pm 2.6\%$

Non-prompt lepton shape modelling Flat rate uncertainty of $\pm 30\%$ used to cover statistical uncertainties

Pileup reweighting Uncertainty in the average expected number of additional interactions per bunch crossing is $\pm 4.6\%$. These new weights are used to generate the uncertainty.

Lepton trigger, ID, and reconstruction efficiency Scaled by adding/subtracting uncertainties from reweighting factor

Jet energy scale Varied by one standard deviation from central value and propagated to MET

Jet energy resolution Momentum varied by one standard deviation from central value and propagated to MET

b **tagging** b tagging scale factor used to reweight the MC is scaled up and down by its uncertainty

Parton distribution functions Taken from LHE event weights

Matrix Element and Patron Shower matching scales Taken from LHE event weights

Strong coupling constant α_S Taken from LHE event weights (not available for leading order MC samples)

MET Modelling of $\not{E}_T$

Statistics Statistics of MC