Chapter 5

Multi-Variate Techniques

In most analyses, several variables can be used to discriminate signal events from the background. A traditional approach is to reject most of the background events by applying cuts on all but one variable. The distribution of the last variable is then used to compute the statistical significance of the measurement. However, this approach does not use the information that is available in the correlations among variables. Multivariate classifiers can exploit such correlations and increase the sensitivity of the analysis. While several implementations of different classifiers and learning algorithms are available, in general, the mode of operation is the same. A training set, which comprises of the input variables and the desired output for each event, is used to optimize the parameters of the classifier. The classifiers output is then used to separate the signal from the backgrounds and to build a distribution that can be used to test the hypothesis under study. In recent years statisticians have found new ways to tune and combine classifiers to further increase the performance of these methods. A package, Toolkit for Multivariate Data Analysis with ROOT (TMVA) [57] is a popular choice at D0 for training and evaluating multivariate discriminants.

All multivariate classification methods must be trained and evaluated on separate, independent samples to ensure an unbiased result. The training should be such that the general trends and correlations of a sample and not the random fluctuations dictate the
an operating point and applying the scale factor to correct the simulation efficiency has a more accurate sensitivity to detector performance and topological correlations. The weights applied to an event with \( p \) number of the jets out of the total of \( n \) jets passing the operating point cut is:

\[
    w_n = \frac{\prod_{i=0}^{p} \text{TRF}_d(P_i, \eta_i) \times \prod_{j=p+1}^{n} \text{TRF}_d(P_j, \eta_j)}{\prod_{i=0}^{n} \text{TRF}_m(P_i, \eta_i) \times \prod_{j=0}^{n-p} \text{TRF}_m(P_j, \eta_j)}
\]  

(4.3)

where, \( \text{TRF}_d \) (\( \text{TRF}_m \)) is the tag rate functions (TRF) for data (MC). If a relatively low operating is chosen, the remainder of the distribution provides useful information that can still be exploited. In this case the weights are applied in a pseudo-continuous manner. That is given jet weight is calculated based on the data and simulation efficiencies at the two adjacent operating points as:

\[
    w_n = \frac{\prod_{i=1}^{p} \text{TRF}_d^b(P_i, \eta_i) - \text{TRF}_d^a(P_i, \eta_i)}{\prod_{i=1}^{n} \text{TRF}_m^b(P_i, \eta_i) - \text{TRF}_m^a(P_i, \eta_i)}
\]  

(4.4)

where, \( \text{TRF}_d^b \) (\( \text{TRF}_d^a \)) is the TRF for the operating point below (above) the jet in question. This is referred to as pseudo-continuous reweighting. In the limit of infinite operating points, this produces a smooth and continuous function to correct the simulation. Since an approximation of this many operating points is not feasible so this method serves as a suitable alternative.
track probability variables, and four variables characterizing track-jets constructed from the set of the tracks selected by JLIP algorithm are the input to one RF. Five more RFs are trained, each having 27 SVT related input variables corresponding to running the SVT with progressively higher quality tracks. These SVT related variables range from those involving track multiplicity, track momentum comparisons, decay lengths, and the dimensions of the track jets built from the selected tracks. These six RFs are then combined in a MLP NN. By replacing the light flavor sample in the training with a c-quark sample a new discriminant known as bc tagger has been built with the aim of separating jets originating from b and c quark hadronization. A third discriminant the bb-tagger is trained to distinguish b-jets from gluon splitting and those from a boson decay. Samples of inclusive, b\bar{b}, and c\bar{c}, taggable QCD MC jets as well as hadronic decaying Z boson MC are used to train these taggers.

4.3.2 Usage

Although the MVA tagger produces output in the form of a continuous distribution, it can only be used in a pseudo-continuous way. The efficiency of the MC does not match that of the data and the correction factors are calculated at several distinct output values thus limiting the user to make requirement relative to these operating points. A jet is said to be tagged if it has an MVA output larger than the operating point of choice. The system D method where (D=8) along with two samples enriched with heavy flavor and muonic jets are used to estimate the b, c, and the light jet efficiency producing tag rate functions (TRFs) for data jets and scale factors to correct the simulated jet tagging efficiency to match these rates for each operating point. TRFs are calculated as function of $P_t$ and $\eta_{det}$.

This information can be used in two ways when producing a simulated flavor enriched sample. One way is to multiply the events weighted by the probability that each jet could be tagged. This preserves the statistics but it has been seen that actually placing a cut
Figure 4.10: (The NN tagger compared to the JLIP, SVT and CSIP taggers shows an efficiency increase of 20-50% for a fake rate of 0.2% and 15% for a fake rate of 4%. (top) For jets of $P_T > 40$ GeV and $0 < \eta < 0.8$ the MVA tagger identifies jets from b-quark hadronization 13% more efficiently while rejecting 50% more of the light jet background (bottom).
Both the NN and MVA taggers use information from two different impact parameter (IP) and one secondary vertex tagger trained on a sample of b-jet (signal) and the light jets (background) such that when evaluated on the properties of a random jet, higher values are assigned to the b-jets while the lower values to the light jets. Traditional use of this information is to place a requirement on the number of jets above an optimally chosen value in order to select a heavy flavor enriched sample.

The two IP taggers used are Jet Lifetime Probability Tagger (JLIP) and Counting Signal Impact Parameter (CSIP) algorithms. JLIP computes the probability of a track originating from a PV given the impact parameter. These probabilities for all the tracks within a jet are combined into the variable known as JLIP Probability. Since b-jets create vertices displaced from the primary interaction point, the JLIP probability for all objects tends towards zero while light jets have a more uniform probability. CSIP calculates the signed impact parameter significance \( S_d = \frac{IP}{\alpha_{IP}} \) for high quality tracks within the jet cone. A jet is tagged if there are at least two high quality tracks with \( S_d/\alpha_{IP} > 3 \) or at-least three with \( S_d/\alpha_{IP} > 2 \). Where \( \alpha \) is the re-normalization parameter.

The Secondary Vertex Tagger (SVT) uses tracks of a chosen quality to reconstruct secondary vertices with the Kalman filter and if one is found in a jet cone, the jet is tagged. The NN tagger uses six SVT variables calculated with tracks having \( S_d > 3 \), two variables from the JLIP tagger, and the one from the CSIP tagger.

### 4.3.1 Multivariate b-Tagging

The overall strategy of the MVA b-tagger is to combine information from the IP taggers in a Random Forest (RF), information from the SVT algorithm in a separate set of RFs and combine these in a MLP NN to exploit the non-linear correlations. RFs were chosen due to their stability in terms of the non-discriminating variables and relatively quick training time. Nine IP tagger based variables, the one’s used in the NN, two additional
most often used by the random forests are the dijet invariant mass, the leading jet $p_T$, $\Delta\phi$ between the leptons and the scaler sum of the $p_T$ from two jets and the two leptons. The dijet invariant mass distribution is shown in the figure and the distribution of other three variables are shown in figure.

All the 20 variables used as the input to the RF training are:

- **Jet Variables:** Following jet related variables were used as the input to the RF discriminant.

  1. dijet invariant mass, both after and before the kinematic fit.
  2. $p_T$ of each jet in the dijet pair both before and after the kinematic fit.
  3. $p_T$ of the dijet system.
  4. $\Delta\phi$ and $\Delta\eta$ between the jets in the dijet pair.
  5. invariant mass and the $p_T$ of the system of all jets.

- **Lepton Variables:**

  1. $p_T$ of the dilepton system.
  2. $\Delta\phi$ and $\Delta\eta$ between the leptons.
  3. collinearity of the leptons in the dilepton system.

- **Global Variables:** These variables, related to both jets and the leptons are used as input to the discriminant.

  1. invariant mass of the $llbb$ system.
  2. $p_T$ of the $llbb$ system.
  3. scaler sum of the lepton and jet $p_T$'s in the $llbb$ system.
  4. $\Delta\phi$ between the dilepton and the dijet system.
  5. cosine of the angle $\theta$ between the beam and the dilepton system, in the rest frame of the initial state.
6.3 Multivariate Analysis Plots

6.3.1 Pre-Tag Input Plots

The pre-tag input distributions for Random Forest (RF) are shown in Figures 6.17, 6.18, 6.19, 6.20, 6.21. These are the distributions which are given as an input to the RF and all of them are perfectly well-modeled.
Figure 6.17: Run 2b pretag sample: (a) dijet invariant mass $M_{WW}$, (b) log, (c) $p_T^1$, (d) log, (e) $p_T^2$, (f) log.
Run 2b1 dimuon

<table>
<thead>
<tr>
<th>inclusive</th>
<th>2j-multijet</th>
<th>2j-pretag</th>
<th>ST</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>95394</td>
<td>4007</td>
<td>3682</td>
<td>84</td>
</tr>
<tr>
<td>all bkg</td>
<td>95228 ± 60</td>
<td>4088 ± 10</td>
<td>3785.8 ± 8.2</td>
<td>98.48 ± 0.67</td>
</tr>
<tr>
<td>Multijet</td>
<td>1412 ± 31</td>
<td>42.81 ± 0.51</td>
<td>11.65 ± 0.26</td>
<td>1.481 ± 0.094</td>
</tr>
<tr>
<td>Zjj</td>
<td>90805 ± 51</td>
<td>3273 ± 10</td>
<td>3068.5 ± 8.0</td>
<td>9.18 ± 0.29</td>
</tr>
<tr>
<td>Zbb</td>
<td>773.7 ± 1.8</td>
<td>210.1 ± 1.5</td>
<td>197.33 ± 0.81</td>
<td>48.59 ± 0.32</td>
</tr>
<tr>
<td>Zcće</td>
<td>2014.0 ± 3.1</td>
<td>459.8 ± 1.5</td>
<td>426.6 ± 1.4</td>
<td>28.56 ± 0.42</td>
</tr>
<tr>
<td>ZZ</td>
<td>45.98 ± 0.21</td>
<td>23.87 ± 0.15</td>
<td>22.66 ± 0.15</td>
<td>1.761 ± 0.038</td>
</tr>
<tr>
<td>WZ</td>
<td>56.51 ± 0.45</td>
<td>25.58 ± 0.38</td>
<td>24.46 ± 0.38</td>
<td>1.022 ± 0.029</td>
</tr>
<tr>
<td>WW</td>
<td>65.50 ± 0.60</td>
<td>7.67 ± 0.42</td>
<td>4.90 ± 0.41</td>
<td>0.169 ± 0.023</td>
</tr>
<tr>
<td>tt</td>
<td>54.62 ± 0.46</td>
<td>45.23 ± 0.44</td>
<td>29.71 ± 0.40</td>
<td>7.73 ± 0.28</td>
</tr>
<tr>
<td>ZH(115)</td>
<td>1.816 ± 0.012</td>
<td>1.344 ± 0.011</td>
<td>1.296 ± 0.011</td>
<td>0.3232 ± 0.0044</td>
</tr>
</tbody>
</table>

Run 2b2-4 dimuon

<table>
<thead>
<tr>
<th>inclusive</th>
<th>2j-multijet</th>
<th>2j-pretag</th>
<th>ST</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>443411</td>
<td>19382</td>
<td>17856</td>
<td>458</td>
</tr>
<tr>
<td>all bkg</td>
<td>439383 ± 318</td>
<td>19101 ± 46</td>
<td>17660 ± 44</td>
<td>429.0 ± 3.2</td>
</tr>
<tr>
<td>Multijet</td>
<td>9456 ± 89</td>
<td>318.7 ± 1.7</td>
<td>91.05 ± 0.90</td>
<td>10.23 ± 0.30</td>
</tr>
<tr>
<td>Zjj</td>
<td>416308 ± 304</td>
<td>15294 ± 45</td>
<td>14369 ± 43</td>
<td>39.4 ± 1.3</td>
</tr>
<tr>
<td>Zbb</td>
<td>3514.3 ± 7.0</td>
<td>952.4 ± 1.0</td>
<td>900.2 ± 3.9</td>
<td>209.9 ± 1.9</td>
</tr>
<tr>
<td>Zcće</td>
<td>9151 ± 16</td>
<td>2089.8 ± 8.3</td>
<td>1943.6 ± 7.9</td>
<td>123.6 ± 2.0</td>
</tr>
<tr>
<td>ZZ</td>
<td>196.41 ± 0.75</td>
<td>105.02 ± 0.54</td>
<td>99.96 ± 0.53</td>
<td>7.57 ± 0.16</td>
</tr>
<tr>
<td>WZ</td>
<td>235.8 ± 1.2</td>
<td>111.80 ± 0.84</td>
<td>106.87 ± 0.82</td>
<td>4.76 ± 0.18</td>
</tr>
<tr>
<td>WW</td>
<td>289.8 ± 2.4</td>
<td>33.49 ± 0.85</td>
<td>20.25 ± 0.66</td>
<td>0.76 ± 0.14</td>
</tr>
<tr>
<td>tt</td>
<td>232.2 ± 2.8</td>
<td>195.4 ± 2.5</td>
<td>129.0 ± 2.0</td>
<td>32.9 ± 1.1</td>
</tr>
<tr>
<td>ZH(115)</td>
<td>7.662 ± 0.046</td>
<td>5.951 ± 0.042</td>
<td>5.736 ± 0.042</td>
<td>1.370 ± 0.021</td>
</tr>
</tbody>
</table>

Table 6.5: Event yields for the dimuon ($\mu\mu$) channel.
Table 6.6: Parameters for the transfer functions. Three $\eta$ regions, $\eta_1$, $\eta_2$ and $\eta_3$, correspond to $0.0 < \eta < 0.8$, $0.8 < \eta < 1.6$, and $1.6 < \eta < 2.5$, respectively.

<table>
<thead>
<tr>
<th></th>
<th>light jet</th>
<th>b-quark jet</th>
<th>b-quark jet with a muon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\eta_1$</td>
<td>$\eta_2$</td>
<td>$\eta_3$</td>
</tr>
<tr>
<td>$p_1$</td>
<td>-0.100</td>
<td>-0.156</td>
<td>-0.184</td>
</tr>
<tr>
<td>$p_2$</td>
<td>2.46</td>
<td>8.07</td>
<td>15.3</td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.0899</td>
<td>0.123</td>
<td>0.131</td>
</tr>
<tr>
<td>$p_4$</td>
<td>3.40</td>
<td>4.31</td>
<td>5.30</td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.235</td>
<td>0.193</td>
<td>0.219</td>
</tr>
<tr>
<td>$p_6$</td>
<td>0.234</td>
<td>0.538</td>
<td>0.430</td>
</tr>
<tr>
<td>$p_7$</td>
<td>0.918</td>
<td>1.42</td>
<td>1.16</td>
</tr>
<tr>
<td>$p_8$</td>
<td>0.0147</td>
<td>0.00557</td>
<td>0.00697</td>
</tr>
</tbody>
</table>

\[ p_1 \times 10^{-3}, p_2 \times 10^{-2} \]

Table 6.7: Muon $p_T$ resolution function coefficients
Figure 6.18: Run 2bpretag sample: post kinematic fit (a) dijet invariant mass $M_{bb}$, (b) log, (c) $p_t^1$, (d) log, (e) $p_t^2$, (f) log.
Figure 6.19: Run 2bpretag sample: (a) $p_T^{b_1}$, (b) log, (c) post kinematic fit boost $p_T$, (d) log, (e) $\Delta\phi(b_1, b_2)$ and (f) $\Delta\eta(b_1, b_2)$. 
Figure 6.20: Run 2bpretag sample: (a) $p_T^Z$, (b) log, (c) $\Delta \phi(\ell_1, \ell_2)$, (d) $\Delta \eta(\ell_1, \ell_2)$, (e) colinearity$(\ell_1, \ell_2)$ and (f) $\Delta \phi(Z, bb)$. 
Figure 6.21: Run 2bpretag sample: (a) $M(\ell\ell bb)$, (b) log, (c) spin-basis variable $\cos(\theta^*)$, (d) $H_T(\ell\ell bb)$, (e) $M(\sum \tilde{J})$, and (f) $p_T(\sum \tilde{J})$. temp
6.3.2 Post-Tag Input Plots

The input plots for the Double Tag (DT) selection are shown in the Figure 6.22, 6.23, 6.24, 6.25, 6.26, 6.27, 6.28, 6.29, 6.30, 6.31.
Figure 6.22: Run 2bdimuon ($\mu\mu$) with DT selection: (a) dijet invariant mass $M_{bb}$, (b) log, (c) $p_T^{b_1}$, (d) log, (e) $p_T^{b_2}$, (f) log.
Figure 6.23: Run 2bdimuon ($\mu\mu$) with DT selection: post kinematic fit (a) dijet invariant mass $M_{bb}$, (b) log, (c) $p_T^{bj}$, (d) log, (e) $p_T^{b2}$, (f) log.
Figure 6.24: Run 2bdimuon (μμ) with DT selection: (a) $p_T^{llb}$, (b) log, (c) post kinematic fit boost $p_T(llb)$, (d) log, (e) $\Delta \phi(b_1, b_2)$ and (f) $\Delta \eta(b_1, b_2)$,
Figure 6.25: Run 2bdimuon (μμ) with DT selection: (a) $p_T^Z$, (b) log, (c) $\Delta \phi(\ell_1, \ell_2)$, (d) $\Delta \eta(\ell_1, \ell_2)$, (e) colinearity($\ell_1, \ell_2$) and (f) $\Delta \phi(Z, bb)$. 
Figure 6.26: Run 2bdimuon (μμ) with DT selection: (a) $M(ℓℓbb)$, (b) log, (c) spin-basis variable $\cos(θ^*)$, (d) $H_T(ℓℓbb)$, (e) $M(\sum j_t)$, and (f) $p_T(\sum j_t)$. temp
Figure 6.27: Run 2b-dimuon ($\mu\mu$) with ST selection: (a) dijet invariant mass $M_{bb}$, (b) log, (c) $p_T^1$, (d) log, (e) $p_T^2$, (f) log.
Figure 6.28: Run 2bdimuon (μμ) with ST selection: post kinematic fit (a) dijet invariant mass $M_{hh}$, (b) log, (c) $p_T^1$, (d) log, (e) $p_T^2$, (f) log.
Figure 6.29: Run 2bdimuon ($\mu\mu$) with ST selection: (a) $p_T^{bb}$, (b) log, (c) post kinematic fit boost $p_T(\ell\ell bb)$, (d) log, (e) $\Delta\phi(b_1, b_2)$ and (f) $\Delta\eta(b_1, b_2)$. 
Figure 6.30: Run 2bdimuon ($\mu\mu$) with ST selection: (a) $p_T$, (b) log, (c) $\Delta\phi(\ell_1, \ell_2)$, (d) $\Delta\eta(\ell_1, \ell_2)$, (e) colinearity($\ell_1, \ell_2$) and (f) $\Delta\phi(Z, bb)$.
Figure 6.31: Run 2bdimuon ($\mu\mu$) with ST selection: (a) $M(\ell\ell bb)$, (b) log, (c) spin-basis variable $\cos(\theta^*)$, (d) $H_T(\ell\ell bb)$, (e) $M(\sum \vec{j}_t)$, and (f) $p_T(\sum \vec{j}_t)$. temp
6.3.3 Pre-Tag RF Output Plots

Double Tag (DT) Random Forest (RF) Output Plots for the Pre-Tag Sample are shown in the Figure 6.32, 6.33 along with the similar plots for the Single Tag (ST) in Figure 6.34, 6.35. These are the distributions we use to set limits on ZH production and in these high-statistics plots, it is clear that the RF output is well modeled.
Figure 6.32: Run 2bdimuon ($\mu\mu$) pretag sample DT-trained random forest output (log scale) for (a) $VZ$, and $m_{H} =$ (b) 100, (c) 105, (d) 110, (e) 115 and (f) 120GeV.
Figure 6.33: Run 2bdimuon ($\mu\mu$) pretag sample DT-trained random forest output (log scale) for $m_H =$ (a) 125, (b) 130, (c) 135, (d) 140, (e) 145 and (f) 150GeV.
Figure 6.34: Run 2bdimuon (\(\mu\mu\)) pretag sample ST-trained random forest output (log scale) for \(m_H =\) (a) 100, (b) 105, (c) 110, (d) 115 and (e) 120GeV.
Figure 6.35: Run 2bdimuon (μμ) pretag sample ST-trained random forest output (log scale) for $m_H =$ (a) 125, (b) 130, (c) 135, (d) 140, (e) 145 and (f) 150 GeV.
6.3.4 Post-Tag RF Output Plots for Every Channel

Double Tag (DT) Random Forest (RF) Output Plots for the Post-Tag Sample are shown in the Figure 6.36, 6.37.
Figure 6.36: Run 2bdimuon ($\mu\mu$) with DT selection: random forest output for (a) $VZ$, and $m_H =$ (b) 100, (c) 105, (d) 110, (e) 115 and (f) 120 GeV.
Figure 6.37: Run 2bdimuon ($\mu\mu$) with DT selection: random forest output for $m_H =$ (a) 125, (b) 130, (c) 135, (d) 140, (e) 145 and (f) 150GeV.
Single Tag (DT) Random Forest (RF) Output Plots for the Post-Tag Sample are shown in the Figure 6.38, 6.39. These are the distributions we use to set limits on ZH production.
Figure 6.38: Run 2bmmumu sample with ST selection: random forest output for (a) $VZ$, and $m_H =$ (b) 100, (c) 105, (d) 110, (e) 115 and (f) 120GeV.
Figure 6.39: Run 2bmmmm sample with ST selection: random forest output for $m_H$ = (a) 125, (b) 130, (c) 135, (d) 140, (e) 145 and (f) 150 GeV.
6.4 Systematic Uncertainties

Most measurements of physical quantities in high energy physics involve both a statistical uncertainty and an additional "systematic" uncertainty. Systematic uncertainties play a key role in the measurement of physical quantities, as they are often of comparable scale to the statistical uncertainties. However, the definition of these two sources of uncertainty in a measurement is in practice not clearly defined, which leads to confusion and in some cases incorrect inferences. A coherent approach to systematic uncertainties is, however, possible.

Statistical uncertainties are the result of stochastic fluctuations arising from the fact that a measurement is based on a finite set of observations. Repeated measurements of the same phenomenon will therefore result in a set of observations that will differ, and the statistical uncertainty is a measure of the range of this variation. By definition, statistical variations between two identical measurements of the same phenomenon are uncorrelated, and we have well-developed theories of statistics that allow us to predict and take account of such uncertainties in measurement theory, in inference and in hypothesis testing. Systematic uncertainties, on the other hand, arise from uncertainties associated with the nature of the measurement apparatus, assumptions made by the experimenter, or the model used to make inferences based on the observed data. Such uncertainties are generally correlated from one measurement to the next, and we have a limited and incomplete theoretical framework in which we can interpret and accommodate these uncertainties in inference or hypothesis testing. Common examples of systematic uncertainty include uncertainties that arise from the calibration of the measurement device, the probability of detection of a given type of interaction (often called the "acceptance" of the detector), and parameters of the model used to make inferences that themselves are not precisely known. The definition of such uncertainties is often ad-hoc in a given measurement, and there are few broadly-accepted techniques to incorporate them into the process of statistical inference.
As our analysis is purely based upon the comparison between the data and the expected background and the signal predictions, the systematic uncertainties are ignored in our interpretation of the data. For example, the JES may not be entirely correct when applied to the $b$–jets, but as long as this error is common between the data and our background model and signal predictions it does not invalidate our results. If a 115 GeV Higgs mass was apparent in our data, we would expect to see a dijet invariant mass peak closer to the 100 GeV thus we would not infer the higgs mass directly from the dijet mass distribution but instead by comparison with signal hypotheses at several values of Higgs mass. The only systematic uncertainties we consider are those which could lead to the disagreement between the data and the predictions, therefore hiding a true signal or creating a false signal.

We can further restrict the types of the systematic errors of concern to us considering the effect of our background normalization procedure. If the effect of an error is to increase or reduce the event yield of all Monte Carlo samples equally, the normalization procedure compensated for it. After normalization, the event yields are insensitive to this type of the error. Therefore we consider only those systematics with the potential to change the shape of the modeled RF output distribution. Most notably we ignore the uncertainty in the integrated luminosity measurement. Our application of the measured luminosity to all the Monte Carlo only serves to ensure that $k_1$ and $k_2$ are close to unity.

It is useful to distinguish between two remaining types of systematic uncertainties of importance to this analysis. The simpler type is the flat systematic, which only effects the event yield of a particular background or signal model. Although a flat systematic does not alter the shape of the RF output distribution for the specific model in question. The shape of the total RF distribution may change due to one background or signal model increasing in yield with respect to others. The canonical example of a flat systematic is the theoretical uncertainty in the cross-section for a particular background process. The other type of systematic uncertainty arises because we use the full shape of the RF
output to obtain results. The shape systematic are applied on per-event basis and can not be reduced to an overall scale applied to an entire background or signal model. Shape systematics arise from the identification and calibration of the simulated jets from the application of $b$–$tagging$ TRF’s and from reweighting applied to improve the accuracy of our Monte Carlo samples, to name a few sources. In the current section, we describe the sources and our handling of all flat and shape systematics considered for this analysis;

**Flat Systematics**

While our normalization procedure allows us to ignore some systematic uncertainties, the resultant event yields remain somewhat uncertain, and those uncertainties have several sources, as the source of error controls how the error should be handled in our limit setting procedure. In the following list, all the flat systematics arising from our background normalization procedure are described:

The uncertainties due to normalization and cross-section are so-called flat systematics which scale all affected events and thereby the corresponding RF distributions by a constant factor. The determination of the normalization constants, i.e. that of the multijet scale factors ($\alpha^{ij}$), that of the channel ($i$) dependent efficiencies ($k_i$), and that of the jet ($j$) dependent cross section factors ($k_j^Z$) is described in Section 6.2.3. The results quoted in Table 6.4 show statistical uncertainties only.

We also assess systematic uncertainties for the normalization procedure. First we repeat the procedure, but allow $k_0^Z$ to vary according to the uncertainty on the inclusive $Z$ cross section. This gives rise to an uncertainty $\sigma^{COR}=8\%$ that is correlated across all channels. We also repeat the normalization procedure for each channel in isolation. The difference between the value of $k_2^Z$ obtained from the combined normalization, and the value obtained from the fit to each channel, $\sigma_1^{IND}$, is assessed as an uncertainty for that channel. In the case of the RunIIb1 $\mu trk$ channel we use instead the statistical precision (5%) of the independent measurement of $k_2^Z$ in that channel. Similarly, we
use a 2% uncertainty for the RunIIb2-4 eeicr channel. Table 6.8 displays the channel independent measurements of $k_L^2$, and the values for $\sigma_{iND}$. A more detailed description of these uncertainties may be found in [75].

<table>
<thead>
<tr>
<th>channel</th>
<th>$k_L^2$</th>
<th>systematic ($\sigma_{iND}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RunIIb1</td>
<td></td>
</tr>
<tr>
<td>CC-CC</td>
<td>0.97 ± 0.03</td>
<td>-3.3%</td>
</tr>
<tr>
<td>CC-EC</td>
<td>1.04 ± 0.04</td>
<td>4.3%</td>
</tr>
<tr>
<td>eeicr</td>
<td>0.94 ± 0.04</td>
<td>-6.1%</td>
</tr>
<tr>
<td>$\mu\mu$</td>
<td>0.97 ± 0.02</td>
<td>-3.4%</td>
</tr>
<tr>
<td>$\mu\mu$trk</td>
<td>0.97 ± 0.05</td>
<td>-2.8% (use 5%)</td>
</tr>
<tr>
<td></td>
<td>RunIIb2-4</td>
<td></td>
</tr>
<tr>
<td>CC-CC</td>
<td>1.00 ± 0.01</td>
<td>-0.52%</td>
</tr>
<tr>
<td>CC-EC</td>
<td>0.98 ± 0.02</td>
<td>-2.2%</td>
</tr>
<tr>
<td>eeicr</td>
<td>1.00 ± 0.02</td>
<td>0.064% (use 2%)</td>
</tr>
<tr>
<td>$\mu\mu$</td>
<td>1.011 ± 0.008</td>
<td>1.0%</td>
</tr>
<tr>
<td>$\mu\mu$trk</td>
<td>1.07 ± 0.02</td>
<td>7.4%</td>
</tr>
</tbody>
</table>

Table 6.8: Measurement of $k_L^2$ by each channel independently

The remaining uncertainties from cross-sections (with corrections described in Section 6.2.1 under the headings $Z/\gamma^* + jets$ Cross-Section and Cross-Sections for other Processes) and are applied to the corresponding MC samples as listed in Tables 6.9 and 6.10.

The normalization and cross-section efficiencies other than $\sigma_{iND}$ are listed in Table 6.9. is listed in Table 6.10. In addition to quoting the uncertainties as a percentage of the individual variable, the overall effect on the background prediction is quantified as a corresponding percent change to the total DT plus ST background. For example, a 20% systematic that affects only 10% of the background has an overall 2% effect.

The PDF uncertainty is estimated by calculating the change in event yield for each of the twenty PDF eigenvectors in CTEQ6M, obtained from the $t\bar{t}$ caf.pdfreweight processor. These twenty uncertainties are then added in quadrature to produce a single flat uncertainty. The difference in the absolute cross section has been removed from this change in event yields, because this effect is already included in the cross section.
Table 6.9: Normalization and cross-section uncertainties common to all leptonic channels, with the uncertainty quoted as a percentage of the individual variable, with the corresponding percent change to the total DT plus ST background (bkg) for the complete data set to which it applies (i.e. RunIIa, RunIIb, or both).

<table>
<thead>
<tr>
<th>Systematic</th>
<th>Uncertainty</th>
<th>bkg</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{\text{COR}}$</td>
<td>8%</td>
<td>1.1%</td>
</tr>
<tr>
<td>$k_2^B$ (RunIIb)</td>
<td>0.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>$k_2^Z$ (CC-EC)</td>
<td>1.69%</td>
<td>1.1%</td>
</tr>
<tr>
<td>$r_B^b$</td>
<td>9%</td>
<td>1.2%</td>
</tr>
<tr>
<td>$r$(EC)</td>
<td>20%</td>
<td>2.6%</td>
</tr>
<tr>
<td>$\gamma^*/Z +HF$ xsec</td>
<td>20%</td>
<td>14.0%</td>
</tr>
<tr>
<td>$t\bar{t}$ xsec</td>
<td>10%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Diboson xsec</td>
<td>7%</td>
<td>0.2%</td>
</tr>
<tr>
<td>$ZH$ xsec</td>
<td>6%</td>
<td></td>
</tr>
</tbody>
</table>

uncertainties. This procedure is repeated independently for each of the following categories: signal (taking the uncertainty from the mass point with the largest change), $\gamma^*/Z +LF$, $Z\gamma^* + c\bar{c}$, $Z\gamma^* + b\bar{b}$, diboson and $t\bar{t}$. These uncertainties are summarized in Table 6.11, and are treated as correlated across all channels and all samples. We have not recalculated these values, but taken them from the estimates obtained for [76].

6.4.1 Jet Reconstruction Uncertainties

We assess systematic changes in the shape of the RF distribution due to uncertainties in the jet energy scale (JES), jet energy resolution (RES) and jet ID efficiency (JETID). In addition to the nominal reconstruction, jets are also reconstructed using values fluctuated at $\pm \sigma$ for JES and RES and $-\sigma$ for jet-ID efficiency to produce five systematic ntuples in addition to the nominal ntuples for each channel. The uncertainties for the vertex confirmation scale factor and the taggability scale factor are now evaluated by adjusting event weights, as described in Section 6.4.2. When setting limits including additional inputs from RunIIa, the RunIIa and RunIIb jet systematics are treated as independent, as the uncertainties are largely statistical.
Table 6.10: Applicability of normalization and cross-section systematic uncertainties to each specific sample, and whether the systematic is independent or correlated between RunIIa and RunIIb.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma^*/Z +\ell\ell$</td>
<td>0.99%</td>
</tr>
<tr>
<td>$Z\gamma^* + bb$</td>
<td>2.36%</td>
</tr>
<tr>
<td>$Z\gamma^* + cc$</td>
<td>1.14%</td>
</tr>
<tr>
<td>Diboson</td>
<td>0.66%</td>
</tr>
<tr>
<td>$tt$</td>
<td>5.9%</td>
</tr>
<tr>
<td>Signal</td>
<td>0.55%</td>
</tr>
</tbody>
</table>

Table 6.11: Uncertainties in total event yield due to PDFs.

The normalization procedure, which fixes the $\gamma^*/Z$ contribution to the data, suppresses the effect on the normalization from these sources of systematic uncertainty. To model this behavior, the systematic ntuples are normalized using the same procedure as the nominal ntuples. The MVA trained on the nominal ntuple is then applied (without retraining) to these systematic ntuples. The resulting final discriminant distributions are different from the nominal distributions and are used in the statistical analysis to quantify these sources of systematic uncertainty.
6.4.2 Systematics Affecting Event Weights

In order to produce an accurate background model, discrete Monte Carlo events are reweighted to apply corrections or to better model certain effects. To assess the systematic uncertainty associated with each of these effects on the shape of the RF distribution, first we propagate the covariance matrix of the model parameters to the error of the event weights. Next, we fluctuate the event weights by ±σ of their errors, to obtain the change in the RF distribution.

- The scale factors for vertex confirmation are now implemented through the use of event weights instead of a random removal procedure. Therefore the corresponding systematic uncertainties are now treated through event weights as well (VCJ).

- The systematic uncertainties for the taggability scale factors (TAGG) are treated through event weights.

- Uncertainties in the b-id scale factors (btagLF, btagHF) are propagated to the event weights, according to the prescription from Section 6.2.1.

- A systematic for the muon trigger (mumu_trig) is assessed by over and under correcting the MC as described in Section 6.2.1.

- The Z-pT reweighting is varied within the statistical uncertainty from the fit (zpt) [77].

- The γ*/Z +jets reweighting in leading jet η and second leading jet η (VJet RW) is varied by 1/2 of the correction.

- Three shape dependent systematics are considered for the ALPGEN + PHYTHIA modeling of V+Jets production. The Alpgen reweighting performed to correct the Mjj distribution is applied at ±1σ from the χ² fit for the best MLM matching point (AlpMLM). The factorization and renormalization parameters in Alpgen+Pythia
are applied at one-half and twice the nominal value ($\text{AlpScale}$). Additionally, we apply a systematic for the underlying event model in Alpgen + Pythia ($\text{AlpUE}$) [78]. The uncertainties are varied independently and event-by-event.

The percentage change in the predicted number of events for each background sample and each shape dependent systematics is displayed in Tables 6.12 and 6.13.

As explained in the previous sections that no significant excess of the data over the expected backgrounds was seen hence the limits for the standard model Higgs boson are set. In the first section we would like to explain the method used to set these limits.

### 6.5 Limit Setting Procedure

Before outlining the limit setting procedure we would like to discuss some terms which will be frequently used in the description.

1. Null Hypothesis ($H_0$): This corresponds to the default or the accepted model. This is also referred to as background only hypothesis as in this only the backgrounds to the model are considered.

2. Test Hypothesis ($H_1$): This corresponds to the alternative model chosen to be tested against the experimental data. This may possibly replace the Null Hypothesis as the accepted model. This model is alternatively known as the signal+background model as this contains the contributions both from the signal and the background. The differences between the Null Hypothesis is known as the parameters of interest.

3. Simple Hypothesis: This is the model in which all the parameters of interest are specified.

4. Compound Hypothesis: In this type of the model all the parameters of interest are not specified including those parameters in which the parameter is said to fall in some specified range. A compound hypothesis can be considered to consisting of a
Table 6.12: Systematic uncertainties for parameters used in the lepton combination for both Run 2b1 and Run 2b2-4 ST samples. The table lists the percentage change in the predicted number of events for each background sample and each shape dependent systematic.
Table 6.13: Systematic uncertainties for parameters used in the lepton combination for both Run 2b1 and Run 2b2-4 DT samples. The table lists the percentage change in the predicted number of events for each background sample and each shape dependent systematic.
set of simple hypothesis corresponding to a set of possible values of the undetermined parameters.

5. Parameters of Interest: As described earlier it is the parameter specifies the differences between the Test and the Null hypothesis. There can be several parameters for a pair of test and null hypothesis but there can only be one parameter of interest.

6. Nuisance Parameter: These are potential candidates for the parameter of interest. These can be shared both by \( H_0 \) and \( H_1 \). The model is said to be a simple model if all of these parameters are known. If these parameters have some range of values or some uncertainty then the model described by them is a compound model.

The final set of events of a counting experiment is governed by the Poisson statistics. The sample consists of the events from the background(B) model, signal+background(S+B) model and the data represented by \( H_0 \), \( H_1 \) and the nature with each being the random sampling of the corresponding parent distribution.

The procedure opted in DØ for limit setting is the modified semi-frequentist approach. The likelihood ratio is calculated for each bin \((i)\) which is defined as:

\[
Q = \frac{P\text{(data}|H_1)}{P\text{(data}|H_0)}
\]

where \( P\text{(data}|H_1) \) is the likelihood that the data is consistent with the test hypothesis, and \( P\text{(data}|H_0) \) represent the likelihood of the data being consistent with the null hypothesis (background only hypothesis). Detailed description of the procedure can be found at [79, 80].

The results so obtained are given in chapter 7.