CDF-9286
version 1
April 7, 2008

A Search for the Standard Model Higgs Boson in the All-Hadronic channel using a Matrix Element Discriminant

Rong-Shyang Lu, Ankush Mitra, Song-Ming Wang
Academia Sinica

Daniel Whiteson
University of California, Irvine

Aart Heijboer, Joseph Kroll
University of Pennsylvania

contact: Aart Heijboer

1 Introduction

This note documents a Matrix Element (ME) based search for the Standard Model (SM) Higgs Boson, where the Higgs is produced in association with a W or Z boson (denoted collectively as V, for 'Vector boson'), and where the H decays into a $b\bar{b}$ pair and the V decays hadronically. See Figure (1.

The advantage of this channel is its large branching fraction. However, the background from QCD multi-jet production is very large. The challenge of this analysis is an accurate prediction of this QCD background.

This channel was previously used in Run-I [1], but this analysis is the first to explore this channel in Run-II.

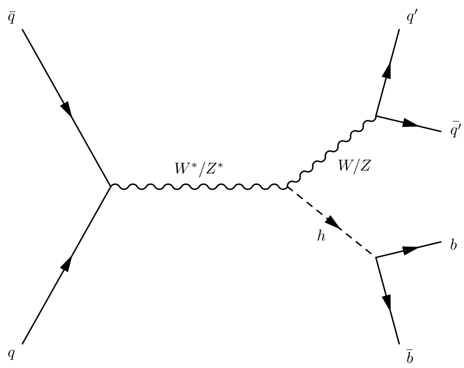

Figure 1: *Feynmann diagram of the* $vh \to q\bar{q}b\bar{b}$ *process*

The analysis presented in this note uses a Matrix Element method to discriminate between signal and background. The analysis shares much of the methodology with a parallel analysis by the same authors that is based on a Boosted Desicion Tree (BDT). Nevertheless, we describe these two analysis in separate notes as there may be subtle differences in implementation. The BDT analysis is described in CDF-9288 [2]. We will refer frequently to this note for more details on certain common elements of the analyses.

2 Data Sample

2.1 trigger

The data are collected using the 'top-multi-jet'-trigger. Originally intended for collecting all-hadronic $t\bar{t}$ events, this trigger requires 4 jets (of 15 GeV at L2) and a large total transverse momentum. There are two main versions of this trigger:

- SUMET_125: requiring a total transverse energy above 125 GeV.
- SUMET_175: requiring a total transverse energy above 175 GeV.

The transverse energy requirement was raised to 175 GeV after the first 430 pb^{-1} of Run-II data taking to reduce trigger bandwidth consumption. A more detailed description of the trigger as well as a detailed study of the trigger acceptance is given in a separate note: CDF-9252 [3].

2.2 offline event selection

The analysis is performed using 2.0 fb^{-1} of data (up to and including period 12 - see [2] for more details). Events are selected from the Stntuples of the SUMET dataset. that have at least four jets, where each jet is required to have:

- Uncorrected transverse energy > 10 GeV
- Corrected (L7) transverse energy > 15 GeV, and
- $|\eta| < 2.4$.

Further requirements we pose on the events are the following:

- The z-coordinate of the primary vertex must be within 60 cm of the origin
- At least one of the four leading jets has to be b-tagged by the SECVTX algorithm. We call events double-tagged if there are exactly two tagged jets among the four leading jets.

- Events are required to be in the 'goodrun-list', version V18 (QCD with Silicon).
- Events with an identified, isolated lepton are vetoed to reduce contributions from $t\bar{t}$ production and to ensure orthogonality with lepton-based Higgs searches.

We exclude events where more than two of the leading four jets are b-tagged. This is done because it greatly simplifies the computation and application of the tag rate function (see section 4.1). Moreover, while we loose some acceptance ($\sim 10\%$ of the Higgs signal has three tags), the discriminant we use does not do a good job of separating these events from the background, since the identification of the Higgs daughters is ambiguous for these events; see figure 2.

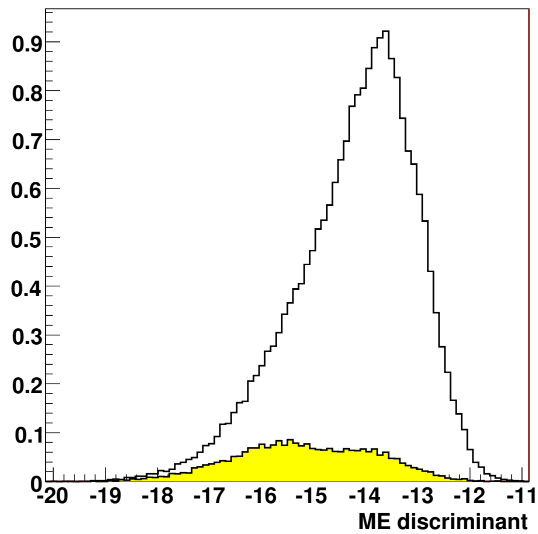

Figure 2: *distribution of the Matrix Element discriminant for VH events with exactly 2 tags (open histogram) and exactly 3 tags (solid). 3-tag events are, on general, more background-like (low value of ME).*

2.3 signal and background yields

The signal monte carlo samples were generated with Pythia and passed through the standard CDF simulation. NLO cross sections taken from [4] are applied to the MC. The trigger acceptance is simulated as explained in [3]. A scale factor is applied to account for the difference in b-tagging efficiency measured in data and MC. This factor is 0.95 ± 0.5 for each tagged

	110	120	130	140	110	120	130	140
cross section	0.216	0.160	0.119	0.090	0.125	0.094	0.071	0.055
Nevt in 2/fb	432.7	319.1	238.6	180.5	249.2	187.1	142.2	109.4
into qq bb	225.3	146.5	85.0	41.9	134.2	88.8	52.4	26.3
4jets, 2tags	13.48	9.86	6.23	3.24	9.26	6.77	4.28	2.21
trigger	3.75	3.03	2.11	1.16	2.80	2.29	1.49	0.86

Table 1: *Cross sections and event yields at various stages of the selection process for WH (left columns) and ZH (right columns).*

.

process	$m_{bb} > 75$ GeV	$m_{bb} < 75$ GeV
data	20899	9961
$t\bar{t}$	319	41
single top	25	9
Wbb	38	5
Z+jets	70	16
WZ	10	1

Table 2: *Event yields in the 'signal' and 'lowbb' regions (see Figure 5 for data and the background processes that are taken into account in the analysis.*

.

jet. More details on the Monte Carlo samples used are given in the note describing the BDT analysis [2].

Table 1 shows the yield of VH events at various stages of the selection process. We finally end up with ~ 5.2 events for $m_H = 120$ GeV. Table 2 shows the 2-tag yields for data and the most important physics backgrounds. See [2] for more information about cross-sections and more processes that have been neglected here, because they yield even less events e.g. $Wc\bar{c}$.

3 Matrix Element Method

In this section, we discuss the discriminating variable we use in this analysis: a likelihood ratio that is constructed from parton-level matrix elements.

3.1 Introduction & Formulas

Matrix Element (ME) methods are a special case of Maximum Likelihood (ML) methods. A central role is played by the likelihood for an event:

$$P(\text{event}|H), \tag{1}$$

where H denotes an hypothesis that we want to test; in this case we want to distinguish between the hypothesis that the event is due to a Higgs decay, and the hypothesis that it is due to a background process.

The ME method computes this likelihood for (in principle) any physics process by starting from the differential cross-section, which can be thought of as a PDF of the parton-level quantities.

The true values of the parton-level momenta are unknown (in contrast to the measured jet energies), which means they must be integrated over to obtain $P(\text{event})$. The observed 'event' is characterized by the momenta of the leading four jets, two of which are tagged and thus associated to parton-level b-quarks. To make the computation tractable, the directions of the partons are assumed to coincide with the directions of the measured jets. This leaves the magnitude of the four parton level momenta as nuisance parameters.

In order to compute the event likelihood, we start with the differential cross-section as function of the two incoming parton momenta (p_1 and p_2) and the four outgoing momenta p_3 through p_6:

$$d\sigma \propto |\mathcal{M}|^2 \times \frac{1}{E_3 E_4 E_5 E_6} \times F(p_1) F(P_2) \times \delta^4(p_1 + p_2 - p_3 - p_4 - p_5 - p_6) \times d^3p_3 d^3p_4 d^3p_5 d^3p_6. \tag{2}$$

Rewriting the differentials of the outgoing momenta in spherical coordinates ($d^3p_i = \rho_i^2 \sin(\theta_i) d\rho_i d\theta_i d\phi_i$, where $\rho_i = |p_i|$) allows us to carry out the integration over the angles while applying δ-functions. This leads to the differential cross section as a function of the magnitudes of the momenta of the four outgoing particles.

$$\frac{d\sigma}{d\rho_3 d\rho_4 d\rho_5 d\rho_6} \propto |\mathcal{M}|^2 \times \frac{\rho_3\rho_4\rho_5\rho_6}{E_3E_4E_5E_6} \times F(p_1^z)F(p_2^z) \times P(P_T), \qquad (3)$$

where the $F(p_i^z)$ are the parton density functions (we use CTEQ5L) and $P(P_T)$ models the transverse components of $p_1 + p_2$; see section 3.1.2.

The differential cross-section, or likelihood, for the *observed* event is now obtained by multiplying with the probabilities $\mathcal{T}$ to measure jet energy E^{jet} when the parton momentum is ρ, and integrating over the four unknown parton momenta ρ_i:

$$\frac{dP(event)}{dE_3^{jet}dE_4^{jet}dE_5^{jet}dE_6^{jet}} = \int \prod_{i=3...6} \mathcal{T}(E_i|\rho_i)\frac{d\sigma}{d\rho_3 d\rho_4 d\rho_5 d\rho_6} d\rho_3 d\rho_4 d\rho_5 \rho_6. \quad (4)$$

We compute the event likelihood by performing a four-dimensional numberical integration over (unknown) true parton momenta.

3.1.1 Jet energy transfer functions

As discussed in the previous section, the computation of the event probability requires $\mathcal{T}(E_{jet}|E_{parton})$; i.e. the PDF of the measured jet energy as a function of the (true) parton energy. We determined the PDF of $E_{jet} - E_{parton}$ for b and non-b jets in a sample of WH MC events for 20 different bins of E_{parton}. these distributions and the fits that are used in the ME computation. For b-jets a pronounced tail is visible on the lower end (because of semileptonic decays).

3.1.2 A-priori net-pt distribution

In the hard scatter described by the matrix element, P_T is strictly conserved. However, initial state radiation will, in general, give the four outgoing jets a small but significant net P_T . To take this into account in the ME calculation, the integral includes a factor $P(P_T)$, which is the PDF of the net P_T of the system. This PDF was determined on WH MC and is shown in figure 4. In the ME calculation this distribution is parameterized by a double Gaussian: about $14\% \times \mathcal{G}(\sigma = 22 \text{ GeV}) + 86\% \times \mathcal{G}(\sigma = 5 \text{ GeV})$.

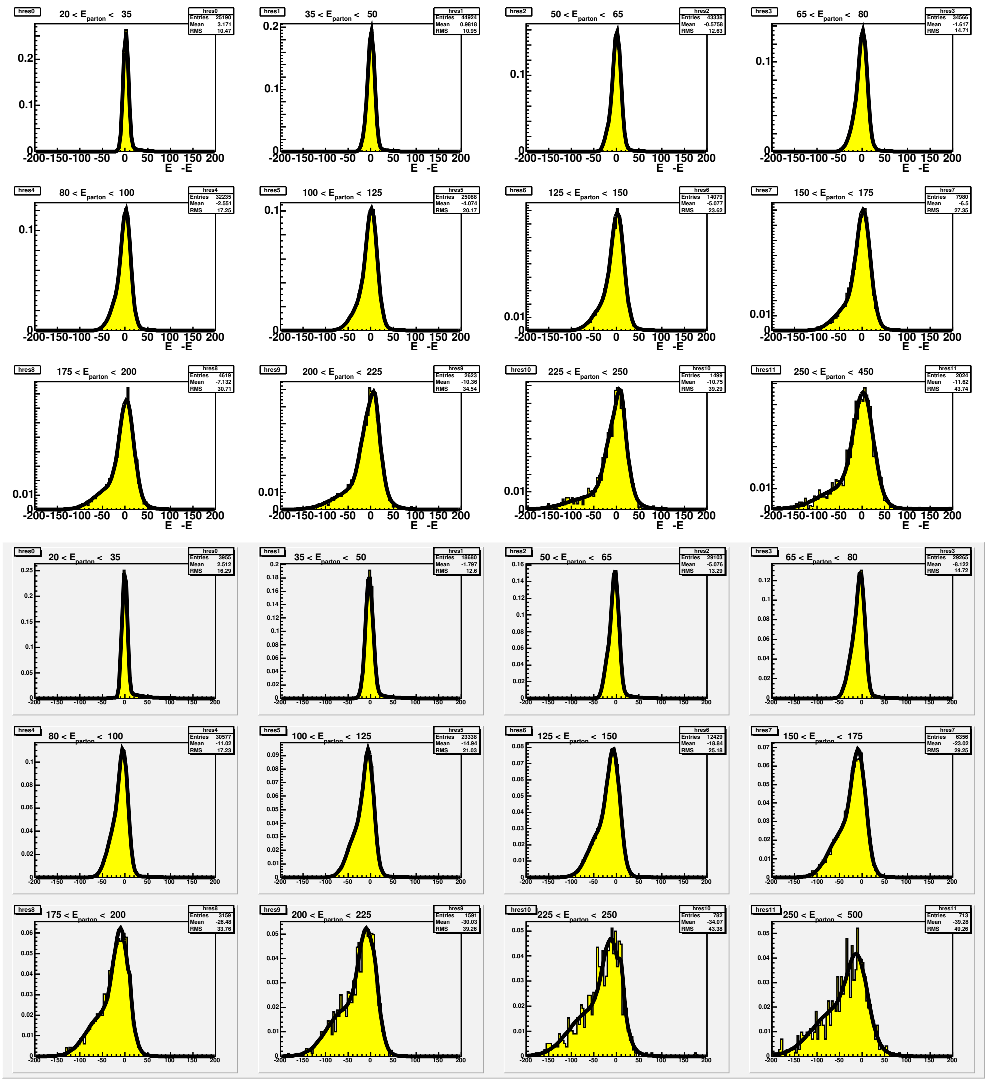

Figure 3: *Distributions of* $E_{\mathrm{jet}} - E_{\mathrm{parton}}$ *obtained from WH MC for non-b jets (top) and b-jets (bottom). The functions are the parameterizations used for the transfer functions in the ME calculation.*

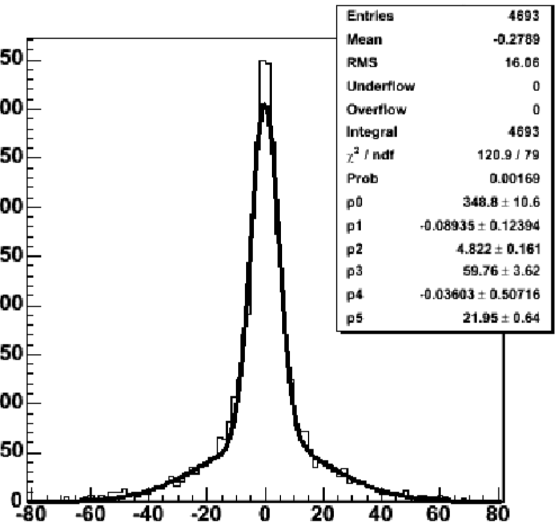

Figure 4: *Distribution of the x-component of the momentum of the VH system from signal Monte Carlo. The double Gaussian fit is used in the ME calculation.*

3.2 Matrix Element Discriminant

We compute matrix element event probabilities (i.e. differential event-cross-sections) for three processes:

- **WH and ZH associated production** for $m_H = 100, 110, 120, 130$, and 140 GeV. The code that computes the matrix element $|\mathcal{M}|^2$ for these processes was taken from Alpgen [5]

- **QCD** $gg \to ggbb$ We use this to obtain an estimate of the probability for background processes. The matrix element for this process is taken from the Madgraph code [6].

It should be noted that the QCD matrix element does not explicitly describe all sources of background; other processes are known to be present (such as $qqbb$ production or $t\bar{t}$). To keep the computation tractable, we only include the most important background process in the ME discriminant. This may, in principle, lead to a degradation of the sensitivity, since the discriminant will not be optimized for distinguishing VH signal from, say, $t\bar{t}$ background.

The quantity we finally use to distinguish signal from background is the log likelihood ratio[1]:

$$Q = \log \left(\frac{P(event|VH) + P(event|ZH)}{P(event|QCD)} \right). \tag{5}$$

Hence, a large value of Q signifies a signal-like events and vice versa.

It was found that for events where the net P_T of the four leading jets is larger than 75 GeV, the value of Q behaves erratically. In these events both numerator and denominator in the above equation are exceedingly small (see Figure 4 to see why), but the ratio may be large; possibly due to numerical effects or other imperfections in the integration. For this reason, we restrict the ME analysis to events with a net $P_T < 75$ GeV, which results in a negligible loss of signal events.

[1]for historical reasons, $P(event|QCD)$ was multiplied by 10^8, so in practice, we use $Q - 8$ as discriminant, which happens to have values between ~ -20 and ~ -10. Of course, this does not effect anything in the analysis.

4 Background Modeling

The main challenge in this analysis is to obtain an accurate model of the background. Of all the processes contributing to the background, QCD $b\bar{b}$ +jets production is dominant by almost two orders of magnitude. Moreover, it is not easy to reliably (we need to know the shape to the few-% level) simulate this background with sufficient statistics. We therefore rely on a data-driven method to model the QCD background. Physics processes producing real vector bosons with $b\bar{b}$ pairs will be modeled by Monte Carlo and are added on top of the data-driven estimate of the QCD background.

4.1 Tag Rate Function

The double-tagged data are modeled by single-tagged events, were each 1-tag event is weighted by the probability of containing an additional tagged jet. This probability is called the Tag Rate Function (TRF). The idea is to measure the TRF using events that do not contain VH signal events; we will use events where m_{qq} is incompatible with the V mass. Since the 1-tag sample is also depleted of signal events (even in for events where m_{qq} is compatible with V), the product of the TRF and the number of 1-tag events yields an estimate of the background that is mostly blind to the VH signal (but see section 4.4).

We have found that the accurate prediction of the shape of the variables we care about, the TRF must be measured as a function of several event variables. We tested a number of different parameterizations on their ability to predict the shapes in a $t\bar{t}$ MC sample (see section 4.4.1). The selected parameterization consists of a 4 parameters: 1) The E_T of the probe jet, 2) The number of 'good tracks' in the probe jet, 3) ΔR between the master jet and the probe jet, and 4) m_{bb}: the invariant mass of the master and probe jets. We denote the vector of these variables by $\mathbf{x}$.

To measure the TRF, events with at least one tag are analyzed. For each of these events, we...

- loop over all the tagged jets in the event. The tagged jet under consideration is called the 'master jet'.
- For each master jet, we loop over each of the other jets that are taggable. This jet other jet is called the 'probe jet'.

- Each combination of master jet and the probe jet forms a candidate double-tagged event: For this pseudo-double-tagged event, we compute the variables $\mathbf{x}$.
- Each pseudo-double-tagged-event enters into the denominator of $TRF(\mathbf{x})$. Iff the probe-jet is actually b-tagged, it also enters in the numerator.

Note that events that have two (real) tags, enter into the numerator twice: each tagged jet will be master once and probe once. We give each of these entries a weight of 1/2 in numerator and denominator.

When measuring and applying the TRF, only the 4 leading jets are considered. more jets may be present, but they are not considered as probe jets. Note that this is consistent with our definition of 'double tagged' events: which requires the two tags to be present in the four leading jets.

4.2 TRF regions

We measure the TRF using several different classes of events, which are defined by the mass of the (candidate) $b\bar{b}$ pair, m_{bb} and the mass of the remaining two jets within the four leading ones, which we call m_{qq}. (So for double-tagged events m_{bb} is simply the invariant mass of the two tagged jets and m_{qq} is the mass of the remaining leading jets.) We measure the TRF in different regions of the (m_{bb} , m_{qq}) plane. The definition of these regions is depicted in figure 5. The default TRF is measured in the region denoted 'tag', which guarantees that the tagging probability is not biased by a possible VH signal. The signal region (denoted 'sig' in the figure) contains the VH events and it is this region we use for the search. The $m_{bb} > 75$ GeV constraint on the signal region was adopted to avoid wasting cpu time computing matrix elements for events that are very background-like anyway.

4.3 Modeling of m_{qq}

As explained before, the TRF is measured in the m_{qq} sidebands. The TRF can, therefore, not depend explicitly on m_{qq} (in contrast to e.g. m_{bb}). As a result, by using the TRF method by itself, we have little hope of modeling m_{qq} correctly in the signal region. In fact, we observe in both data and $t\bar{t}$ MC that the m_{qq} background prediction very closely follows the shape of m_{qq}

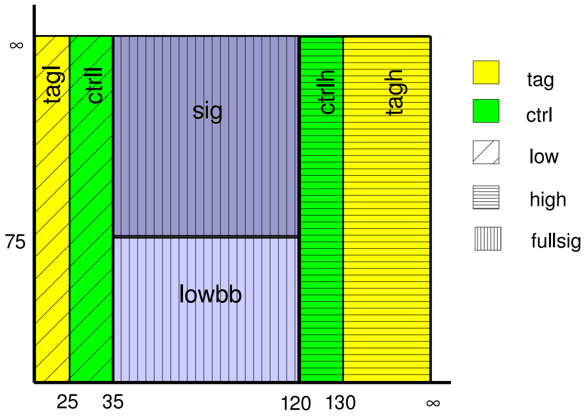

Figure 5: *Depiction of the defined regions in m_{bb} ,m_{qq} space we use for measuring the tag rate function.*

in the 1-tag sample. In other words: as far as the m_{qq} shape is concerned, the TRF method normalizes the 1-tag data to give the 2-tag prediction, but it cannot predict the shape of the 2-tag background. In the data, the TRF prediction happens to model the 2-tag shape reasonably well; this is due to the (perhaps sheer lucky) fact that the 1-tag and 2-tag shapes differ only slightly. However, we would like to endow the background prediction method with some mechanism to model the m_{qq} shape rather than relying on this (coincidental) agreement.

Therefore, we correct the background prediction using a reweighing-function that is derived from the difference between background prediction and data in the region $m_{bb} < 75$GeV (we don't use this region in the rest of the analysis). Before extracting the correction curve from the $m_{bb} < 75$, the background prediction in that region corrected by adding the physics background from MC. The yields of these background to the $m_{bb} < 75$ GeV region were shown in Table 2; just like in the $m_{bb} > 75$, they are very small compared to the number of events in data.

Figure 6 shows relative shape of the residuals in the signal region and in the region used for the reweighting. The curve showed in the latter is a 2nd degree polynomial resulting from a fit to the residuals. This is the function that is used for the reweighting. The shapes are similar, which implies that the reweighting indeed corrects for the (already small) initial discrepancy in the signal region. The resulting m_{qq} distribution in the signal region after reweighting can be seen in Figure 12.

A systematic background shape uncertainty is obtained by also deriving a reweighting curve in a high m_{bb} region ($160 < m_{bb} < 200$ GeV). The difference in the discriminant shape given by the two m_{qq} reweighting functions is the shape uncertainty.

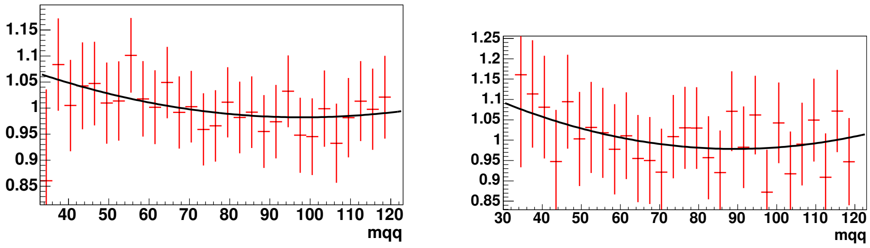

Figure 6: *Comparison between 2-tag data and the background prediction in signal region* $35 > m_{qq} > 120, m_{bb} > 75 GeV$ *(left) and for events with* $35 > m_{qq} > 120, m_{bb} < 75$*. The 2nd degree polynomial fitted to the right figure is used to correct the* m_{qq} *modeling in the left figure.*

4.3.1 Contributions of non-QCD backgrounds

By measuring the TRF in the m_{qq} sidebands, we ensure that the TRF does not include events where the non-tagged jets originate from a real W/Z. To account for such events in the prediction of the 2-tag data, we determine their contribution from Monte Carlo.

However, events from $t\bar{t}$ require special treatment, because: 1) it is by far the dominant background containing real W's and b-quarks, and 2) The high jet multiplicity in hadronic $t\bar{t}$ events results in events where the two non-tagged jets form a mass that lies in the m_{qq} sidebands for a fraction of the events. Hence, a measurement of the TRF in the m_{qq} sidebands will include some of the $t\bar{t}$ signal and will thus over-estimate the tagging probability compared to a TRF that would be measured on pure QCD events.

In order to obtain a TRF that describes the tag-probability for QCD events, we subtract the contribution of $t\bar{t}$ from both numerator and denominator of the TRF. This contribution is taken from the $t\bar{t}$ Monte Carlo. Hence:

$$F_{QCD}(\mathbf{x}) = \frac{N_{tag}^{data}(\mathbf{x}) - N_{tag}^{t\bar{t}}(\mathbf{x})}{N_{tagable}^{data}(\mathbf{x}) - N_{tagable}^{t\bar{t}}(\mathbf{x})}. \tag{6}$$

F_{QCD} is the tag probability of QCD events, with a small contributions from e.g. Z+jets (see section 5.2.2). The number of 1-tag QCD events in the signal region is obtained from data by subtracting the 1-tag $t\bar{t}$ MC. The number of 2-tag QCD events is then obtained by multiplying F_{QCD}. Finally, we add the 2-tag $t\bar{t}$ MC and other 2-tag MC due to single top, $Wb\bar{b}$ and WZ to to arrive at the predicted number of 2-tag events.

$$N_{2tag} = F_{QCD} \times (N_{1tag}^{data} - N_{1tag}^{t\bar{t}}) + N_{2tag}^{t\bar{t}} + N_{2tag}^{otherMC} \tag{7}$$

4.4 Higgs signal influencing the background prediction

The TRF method is based on the assumption that the relative contribution of the VH signal to the 1-tag data is small; i.e. that the VH contribution to the 1-tag sample is very small compared to its relative contribution to the 2-tag sample.

If, however, the VH signal has a sizable contribution to the 1-tag sample, the sensitivity of the search will be reduced. This can be understood by considering the hypothetical (extreme) case where the tag-rate for VH events is equal the the tag-rate for QCD events: $N_{VH}^{1tag}/N_{VH}^{2tag} = N_{QCD}^{1tag}/N_{QCD}^{2tag}$. In this case, any Higgs signal in the 2-tag data, would be mimicked exactly by the 1-tag data and, consequently, the 2-tag-background-prediction ($= TRF \times N_{VH+QCD}^{1tag}$) would always match the observed data, regardless of the VH cross-section. In other words: we would have no possibility to see (or exclude) the signal since any signal would increase our background estimate just as much as it would increase the data. Fortunately, we are not in this regime. However, the contribution of the VH signal to the 1-tag sample is not negligible and we suffer a loss of sensitivity because of this effect.

We take this effect into account by fitting for $N_{VH}^{2tag} - TRF \times N_{VH}^{1tag}$, where the first term is just the number of 2-tag signal events we would normally fit for, and the second terms accounts for the fact that we are insensitive to a fraction of the events because they also add to our background model. The second term amounts to an effective decrease in the signal of about 20% in the high-discriminant region; see Figure 7.

4.4.1 Checks of the TRF in $t\bar{t}$ monte-carlo

In order to check the accuracy of the TRF method, we use it to predict the discriminant shapes in $t\bar{t}$ MC. The reason is that we have a large sample

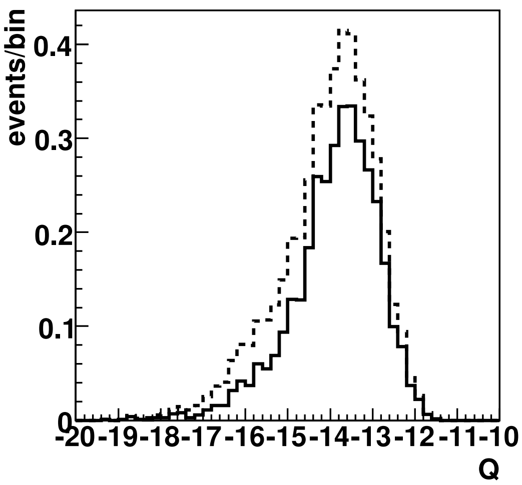

Figure 7: *Effect of the the correction to the signal template for the contribution of the 1-tag VH signal to the background model.*

available, in contrast to QCD MC. While this sample has different properties than the QCD background, we still would expect the TRF method to correctly predict the discriminant shape. Hence, we measure the TRF in the ttbar MC and then apply it to the same sample: any observed discrepancy points to a deficiency of the TRF method to model the 2-tag data. We have no reason to suspect that such a deficiency would occur preferentially in the $t\bar{t}$ MC rather than the data, we will use such discrepancies as one of the ways to estimate the systematic error on the background shape prediction.

As explained before (section 4.3, the TRF cannot (be expected to) model the shape of m_{qq} well. This is particularly important in the $t\bar{t}$ MC. The two-tag-sample (where both b-jets are correctly identified) shows a pronounced W-peak, which is less prominent in the 1-tag sample. Since this effect is much larger in $t\bar{t}$ MC than it is in data, we reweight the events so that the m_{qq} shape agrees between the 2-tag events and the 1-tag-based TRF prediction. This is the only aspect where the treatment of the $t\bar{t}$ MC differs from what we do for real data (where we also reweight m_{qq} , but with a function derived from another region).

In this cross-check, the TRF is measured in in the signal region: the same region where it is applied. It is thus used to predict the same data that is were use to measre the TRF. Hence, any discrepancy that is observed, is due ot the inability of the of the TRF, which is parameterized as a function of only four variables, to fully capture all the aspects of the data that are needed to provide an accurate prediction of the shape of the 2-tag data.

The check on $t\bar{t}$ MC was used to select a TRF parametrization that predicts both the ME and BDT discriminants well. Unfortunately, there is no single ($\leq$ 4-dimensional) TRF parameterization that yields a perfect background prediction shape for every variable. The TRF that was finally adopted is a function of $\mathbf{x} = (\mathbf{E_T}, \mathbf{\Delta R}, \mathbf{m_{bb}}, \mathbf{N_{trks}})$. This parameterization does a good job of predicting the shape of the discriminant variables (both ME and BDT) and models most other variables reasonably well; see Figures 8 and 9. See [2] for a description of the event shape variables sphericity, aplanarity and centrality; like many of the variables shown they are not explicitly used in the computation of the ME discriminant, but we show them anyway.

Throughout this note, the background in such plots is normalized to the data; this reflects the fact that the normalization of the bg-model will be left free in the final fit; the overwhelming number of background events guarantees that the final fit will have no trouble to accurately determine the normalization.

The (small) remaining shape discrepancy in the discriminant distributions (bottom plot in Figure 9) is used as a measure of the systematic error to the background shape prediction due to to the deficiency of the (finite-dimensional) TRF to fully model the shape of the 2-tag sample.

4.5 Checks of the background modeling in the data

VH events with high jet multiplicity are somewhat rare (40(50)% of signal(data) events have $>$ 4 jets). More importantly, signal events in this category are harder to distinguish from background as the identification of the two leading jets with the W or Z daughter may be wrong. Therefore, events with high jet multiplicity offer a 'control-region' where we can test the background model while being relatively insensitive to a possible Higgs contribution. The observed data are compared with the background prediction in Figure 10. The data are modeled well in this region. Naturally, statistics in this sub-sample are reduced compared to the full sample, so we can not use this cross-check to put strong constraints on a possible mismodeling of the background in the full sample. In stead, we conclude that this cross-check shows no reason to suspect the modeling of the background in the full sample would be flawed.

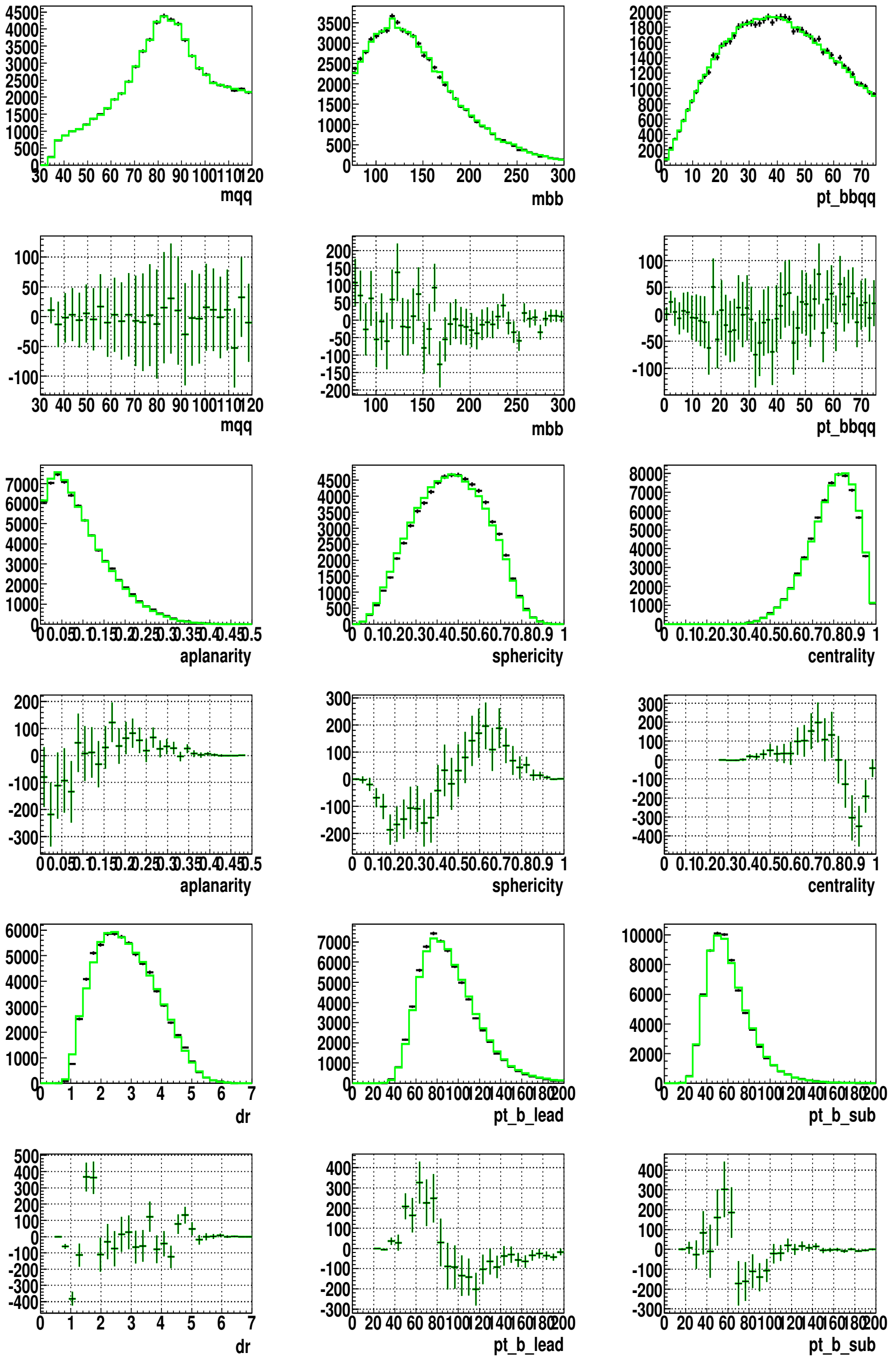

Figure 8: *Comparison of observed 2-tag data in the $t\bar{t}$ MC sample with the shape prediction derived from the 1-tag $t\bar{t}$ MC using the TRF method.(variables such as* pt_u_sub, *denote the P_T of the sub-leading untagged jet.)*

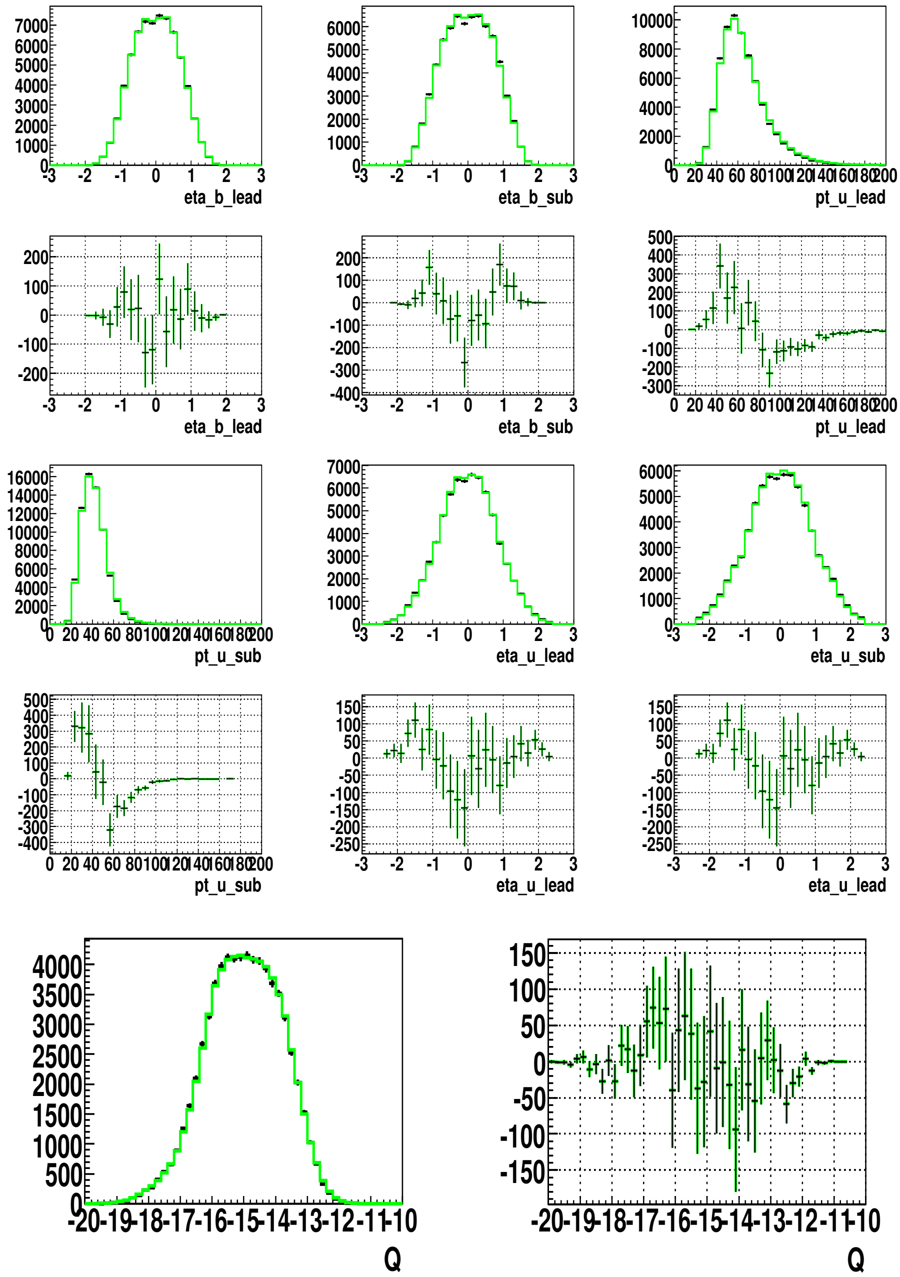

Figure 9: *ttbar cross-checks; continuation of previous figure. The final two plots show the ME discriminant Q for a test-mass of m_H =120 GeV.*

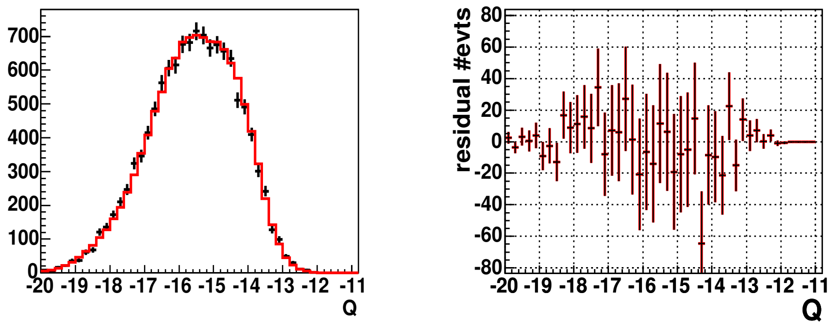

Figure 10: *Comparison of the ME discriminant distribution between data and background prediction in the signal region for events with at least 5 jets. The discriminant is for a test-mass of 120 GeV.*

As another cross check, we use the TRF method to predict the discriminant shape in a different sideband of m_{qq} than the one we use measure the TRF. The TRF is measured in the 'tag' regions (see Figure 5). In Figure 11 it is used to predict the discriminant shape in the $m_{bb} > 75$ GeV part of the 'ctrl' region. The data are modeled well in this region. However, having an m_{qq} incompatible with M_V, these events have a very background-like value of the discriminant. This check therefore does not tell use much about the modeling of the background in the high-Q region. Nevertheless, the correct modeling in this region, reinforces our confidence in background modeling method.

4.6 Comparison of data and background prediction

As a final step in the background modeling checks, we compare distributions of a number of variables (but not the final discriminant yet) with their background predictions. Figures 12 and 13 show these distributions. We conclude from these figures that most variables are modeled well, although there are small shape differences which are clearly not due to statistical fluctuations, but due to few-percent-level mismodeling of the the shape.

It is interesting to note that many of these shape discrepancies are (at least qualitatively) similar to those observed in the $t\bar{t}$ MC shown in Figures 8 and

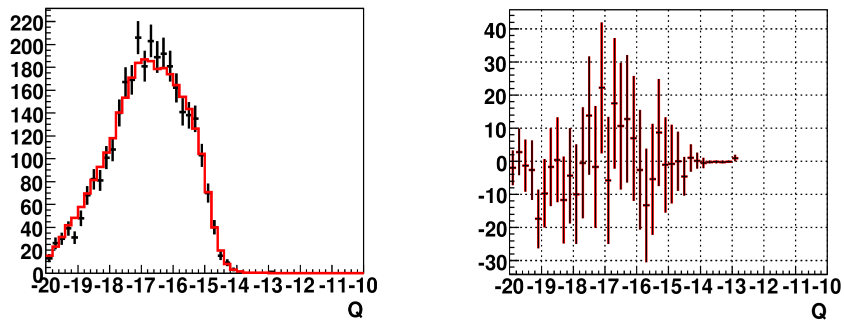

Figure 11: *Comparison of the ME discriminant distribution between data and background prediction in a sideband of* m_{qq} *(the 'ctrl' region), with* $m_{bb} > 75$ *GeV. The discriminant is for a test-mass of 120 GeV.*

9. (Exeptions are the $P_T(jet)$ distributions whose prediction in $t\bar{t}$ MC is too hard, whereas it's slightly too soft in data.) In the $t\bar{t}$ sample, we saw similar residual shape differences as we see in the data, but we still predict the discriminant shape with good accuracy. Moreover, we assigned a systematic for the remaining mismodeling of the discriminant in the $t\bar{t}$ MC. Since the data show similar shape residuals, we can be confident that also in data, the discriminant shape will be largely unaffected and that the systematic on the discriminant shape derived from the $t\bar{t}$ MC will cover any mis-predictions in data as well.

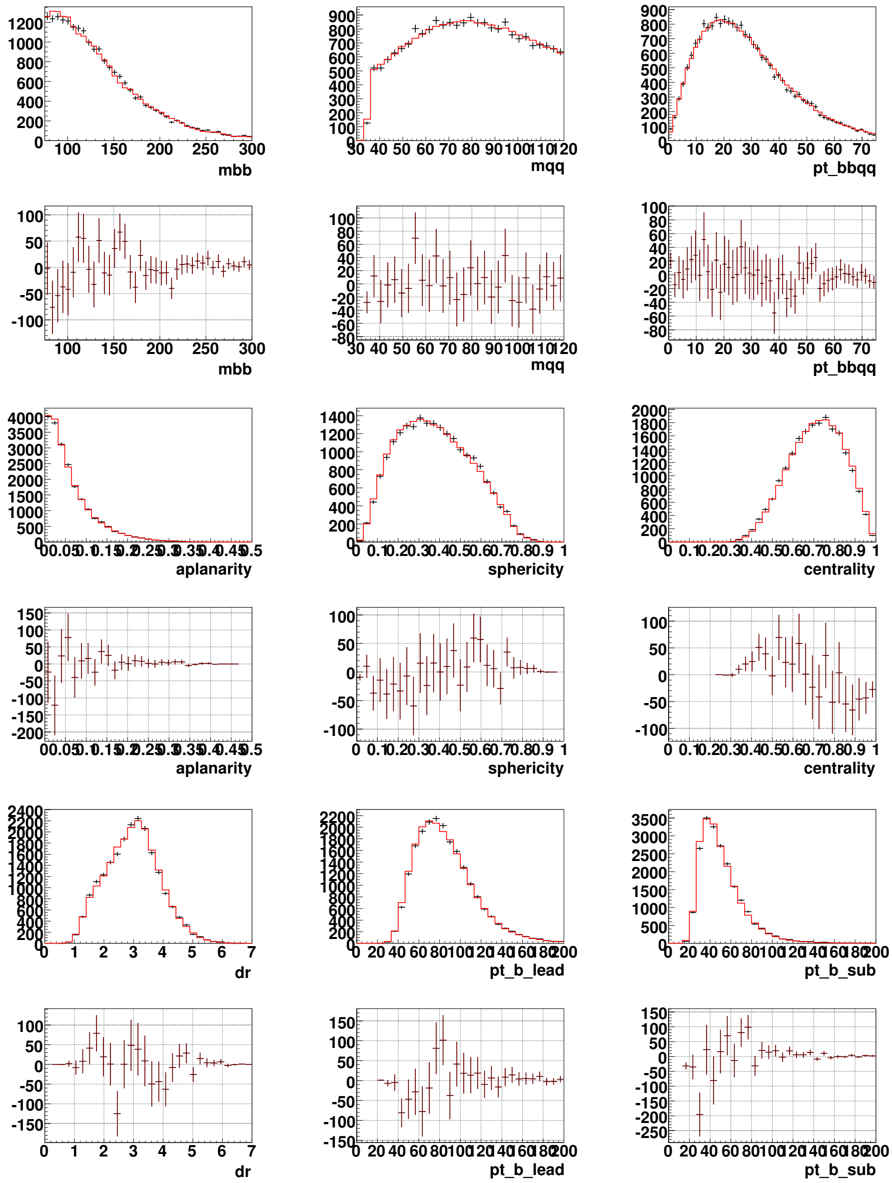

Figure 12: *Comparison of data and background prediction for a number of variables. For each variable, the background prediction (histogram) and data are shown (points with error bars). For each variable the difference between data and residual is shown also.*

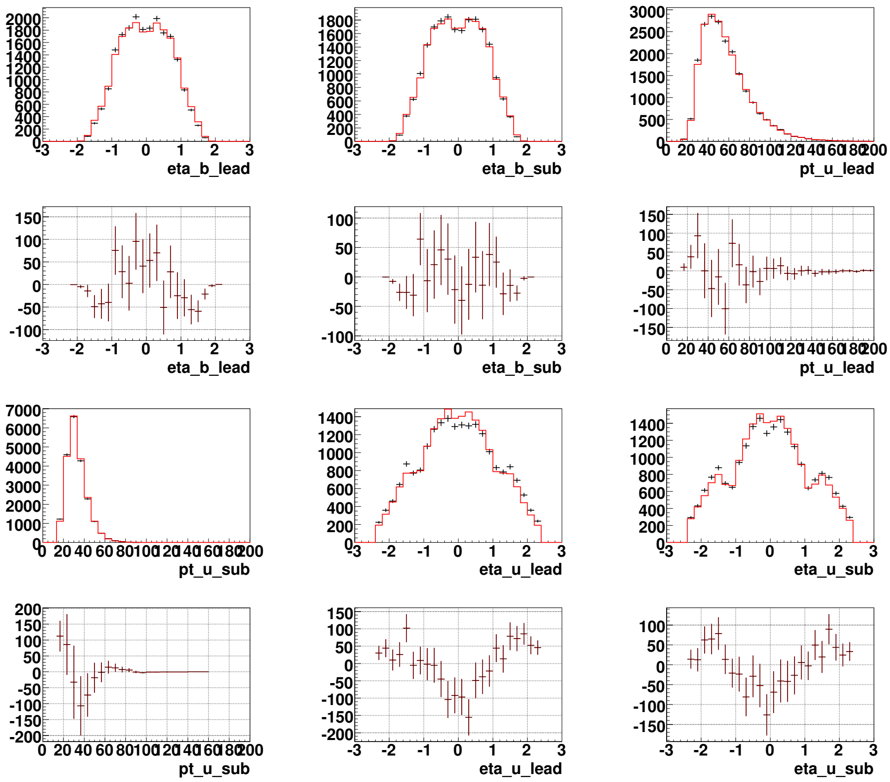

Figure 13: *Comparison of data and background prediction for some more variables.*

5 Fitting and Limits

The data are fit to a simple model, which predicts the contents of the ith bin in a histogram of the the matrix element discriminant at

$$N_i = \beta \times N_i^{bg} + \mu \times N_i^{sig}, \tag{8}$$

where N_i^{bg} is the background model that is constructed using the TRF method as explained in section 4.1, with the corrections for $t\bar{t}$ and other MC as described in section 4.3.1. β is a normalization factor for the total background prediction (note that we could have floated e.g. $t\bar{t}$ separately, but we opted for this simple fit, since having more floating components is not expected to improve the sensitivity). N_i^{sig} is the expected number of signal events for the SM VH cross-section, after the correction described in section 4.4. Finally, μ is the scaled VH cross section $\mu = \sigma/\sigma_{SM}$. When setting limits, we scan over the value of μ; in each individual fit it is set to the hypothesis currently under test. Hence β is the only variable that actually floats in the fits.

The fit for β maximizes the Poisson likelihood of the data. This is done for $\mu = 0$ and $\mu = \mu_{test}$ and the outcome of the experiment is described by the ratio of the Poisson likelihoods[2]:

$$\Lambda = \log\left(\frac{P(bins|\mu = \mu_{test})}{P(bins|\mu = 0)}\right). \tag{9}$$

5.1 Limit setting

Limits on the VH cross section are determined by generating pseudo experiments (PE) to produce distributions of Λ for each hypothesis of the true value of μ. All systematics on the signal and background shapes are assumed to be symmetric; i.e. upward- and downward fluctuations are applied to the pseudo-experiments with equal probability. In the generation of the PE's, each of the systematic uncertainties is multiplied by a normally-distributed random number. The resulting shape is added to the default background shape before the PE is finally generated by throwing Poisson numbers from the (now varied) mean number of events in each bin. This results in distributions for the test-statistic Λ for the background-only case and for the VH

[2]This Λ is more commonly denoted as Q, but we reserve that symbol for the ME discriminant.

cross-section that is currently being tested for. From these distributions, the confidence level of a given measurement (i.e. a given value of Λ) is determined. We compute confidence levels according to the so-called CL_s method [7].

5.2 Systematics

In the following we list all the systematic uncertainties included in the generation of the PE's

5.2.1 Uncertainties on the acceptance

The following systematics for the signal acceptance are included.

- ISR/FSR: From MC generated with less and more initial and final state radiation, shape differences are obtained for the signal templates.
- Jet Energy Scale. In MC, the JES is varied up and down according to the prescription of the the jet energy group.
- PDF's: The uncertainties on the PDF's are taken into account following the prescription given on the CDF joint physics page..
- Trigger acceptance: As detailed in [3], a 7% uncertainty in the overall acceptance of the trigger is assigned.
- Luminosity uncertainty is 6%.
- The B-tagging scale factor uncertainty of 5% per tagged jet leads to a 10% uncertainty in the overall acceptance due to uncertainty in the tagging efficiency.

For the first three items, we model the full shape systematic (which also includes the uncertainty on the overall acceptance). The relative shape variations were derived for a discriminant for a 120 GeV higgs signal and they are applied to all mass points. They are shown in Figure 14. In the region of sensitivity to a VH signal (i.e. roughly between Q=-14 and Q=-12), the dominant uncertainty on the signal is due to to the jet energy scale uncertainty.

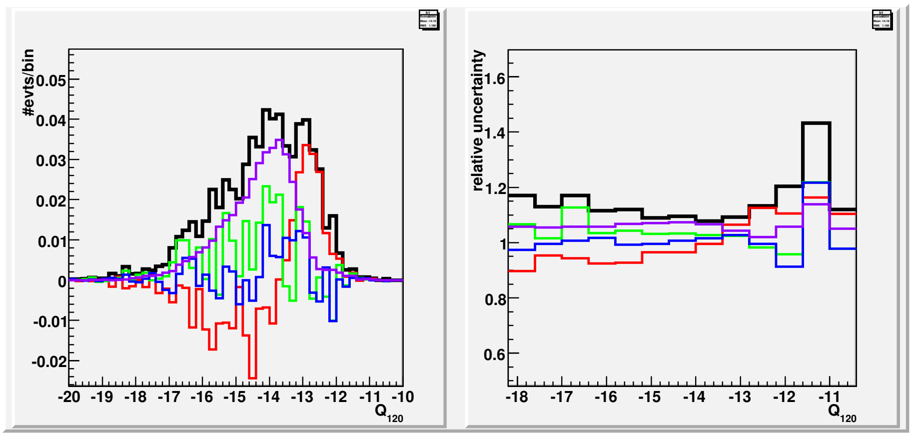

Figure 14: *Systematic uncertainties on the signal shape and acceptance in terms of absolute number of events (for $\sigma_{VH} = \sigma_{SM}$) (left) and as the relative uncertainty on the signal (right). The curves are: combined(black), JES(red), PDF(purple), green(ISR), blue(FSR).*

5.2.2 Physics backgrounds cross-section systematics

- The $t\bar{t}$ cross section uncertainty is taken to be 12%, which is the theory error on the predicted cross-section [8].

- Contributions to the 2-tag sample from single top, $Wb\bar{b}$ and ZW are added added to the background shape prediction. Their contribution is collectively varied by 50% to account for possible uncertainties in the cross-sections.

- Z+jets, where the Z decays into $b\bar{b}$ (or possibly $c\bar{c}$) to form the VH candidate deserves some special attention. Because the mass of the two light jets forming the V candidate is not confined to the signal region, Z+jets also contributes to the m_{qq} sidebands where we measure the TRF. Hence, this signal will be modeled correctly by the TRF, provided its relative contribution to the data (compared to QCD) is not very different in the m_{qq} sidebands than it is in the signal region. Figure 15(right) compares the m_{qq} in the Z+jets events to the (QCD) data. The relative contribution of Z+jets is roughly equal in the signal and sideband regions. Hence, we may expect most of the Z + jets contribution to be modeled correctly by the TRF method. However, Z + jets is slightly more peaked in the signal region ($35 > m_{qq} > 120$)

compared to the QCD. To account for possible mis-modeling due to this discrepancy, we include 30% of the $Z+\text{jets}$ signal as a systematic uncertainty on the background shape. Note that the shape of $Z+\text{jets}$ is very background-like, which makes that this systematic will have only a very small effect on the background shape.

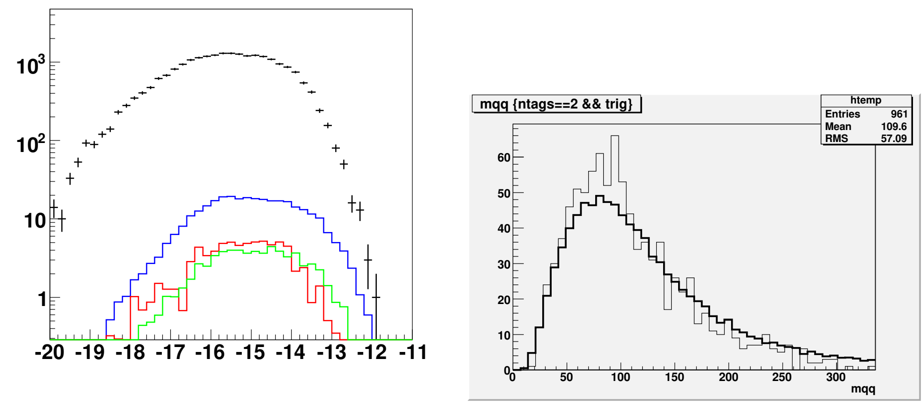

Figure 15: **left:** *Shape of the matrix element discriminant Q for a test-mass of $m_H = 120$ GeV for data (black), $t\bar{t}$ (blue), $Z+\text{jets}$ (red), other MC's (green).* **right:** *Comparison of m_{qq} distribution between data (thick) and $Z+\text{jets}$ MC (thin).*

5.2.3 Background shape systematics

The following uncertainties on the background shape are included:

1. TRF interpolation. The uncertainty associated with measuring the TRF in the m_{qq} sidebands and applying it in the signal region is estimated by using an alternative sideband (called 'ctrl' in Figure 5). The ME discriminant shape difference that results from the variation is applied as a systematic uncertainty.

2. m_{qq} modeling. As explained in section 4.3, we assign an uncertainty for the shape modeling of m_{qq} by using an alternative region to derive the m_{qq} reweighting function.

3. TRF modeling: As explained in section 4.4.1, we run the background modeling procedure on $t\bar{t}$ MC and assign the observed shape difference between 2-tag MC and prediction as a systematic. This uncertainty accounts for the possible inadequacy of the 4-dim TRF to model all the aspects of the data sufficiently to predict the shape of the discriminant.

Figure 5.2.3 shows the three uncertainties on the background shape in terms of absolute number of events. (For the 'TRF interpolation' systematic derived from the $t\bar{t}$ MC, the relative discrepancy is applied to the background prediction in data to arrive at this result).

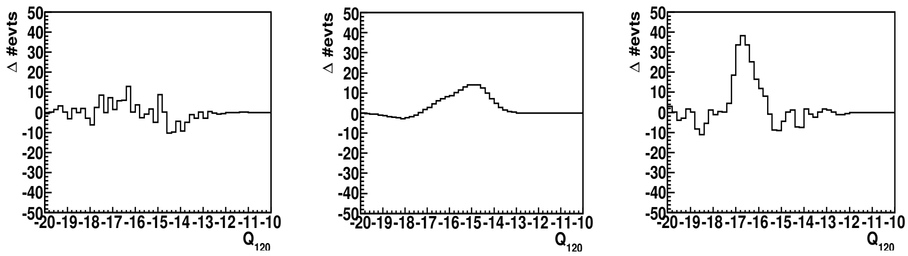

Figure 16: *Systematic uncertainties assigned to the background shape modeling, due to (from left to right) TRF interpolation uncertainty , m_{qq} modeling uncertainty , and TRF modeling uncertainty.*

Rather than fitting these histograms with a function to describe the shape, the histograms themselves are used to vary the background. In the case of 1) and 2) the histogram is a fairly smooth curve to start with. For 3), there is a statistical component to the bins. However, the statistical fluctuations are relatively small, since we have much more $t\bar{t}$ MC events than data events; to remove some of the bin-to-bin fluctuation, the histogram has been smoothed (with ROOT's TH1::Smooth() function).

In order to get a feeling for which values of the ME discriminant are most important for the sensitivity, Figure 17 shows the number of signal events, divided by the square root of the background. From this plot, we conclude that the most important region is roughly $-14 > Q > -12$. Looking at Figure 5.2.3, we see that in this region, the background shape systematics amounts to ~ 10 events per bin. This can be compared with to the signal yield shown in figure 7, which shows that, for a SM VH cross-section, we expect ~ 0.2 events in each bin.

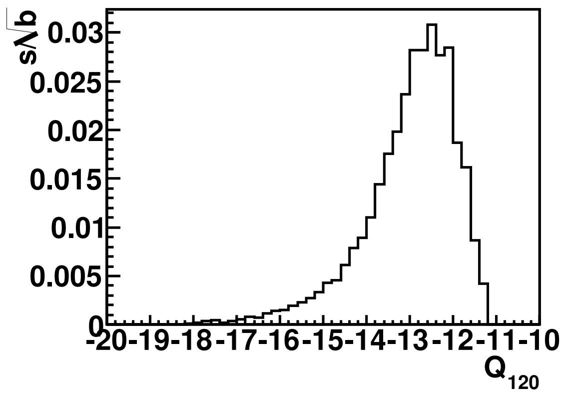

Figure 17: signal/$\sqrt(\text{data})$ *for ME discriminant Q with a test-mass of 120 GeV. This quantity shows how the different regions of the discriminant contribute to the sensitivity.*

;

m_H	$\pm 1\sigma$	$\pm 2\sigma$	median	observed
100	20.7 41.6	14.6 58.0	28.8	27.2
110	23.6 48.7	17.2 68.3	33.6	36.0
120	27.8 56.7	20.2 79.5	39.6	38.0
130	50.4 104.5	35.2 144.5	72.9	66.1
140	82.7 167.4	56.0 235.1	114.5	105.3
150	XXX XXX	XXX XXX	273.5	264.2

Table 3: *Expected and observed limits*

.

6 Results

Figure 6 finally shows the discriminant distribution for the different m_H test masses (from 100 to 140) and the background prediction. In general, the data agree well with the background prediction. The residuals for a test mass of 100 GeV show some structure which may be non-statistical, but the shape does not resemble a VH signal.

From the data, we extract limit on the VH cross section, expressed as a multiple of the SM cross-section. The result is shown in Figure 19, along with the median expected limit and the range of limits that contains 68% of the generated pseudo-experiments.

The observed limit is very close to the median expected limit. This is consistent with Figure 6(right), which shows that the residuals are consistent with zero and do not resemble the expected signal shape. We checked this a bit more quantitatively by doing a simple χ^2 fit of the normalization of the signal shape to the residuals: for each mass point the fitted signal size is consistent with zero to well within 1σ. Some more sanitiy checks on the observed limit will be performed before blessing.

7 Conclusion

We have performed a search for the associated production of the Standard Model Higgs boson. Currently, the analysis is not sensitive to detect a low-mass Higgs signal unless its cross-section would be a factor of ~ 30 larger that what is predicted by the SM.

The data agree (very) well with the data-driven background model and thus show no sign of a large VH-like signal. We set a limit on the cross-section of $\sim 40 \times \sigma_{\mathrm{SM}}$ for $m_H = 120$ GeV.

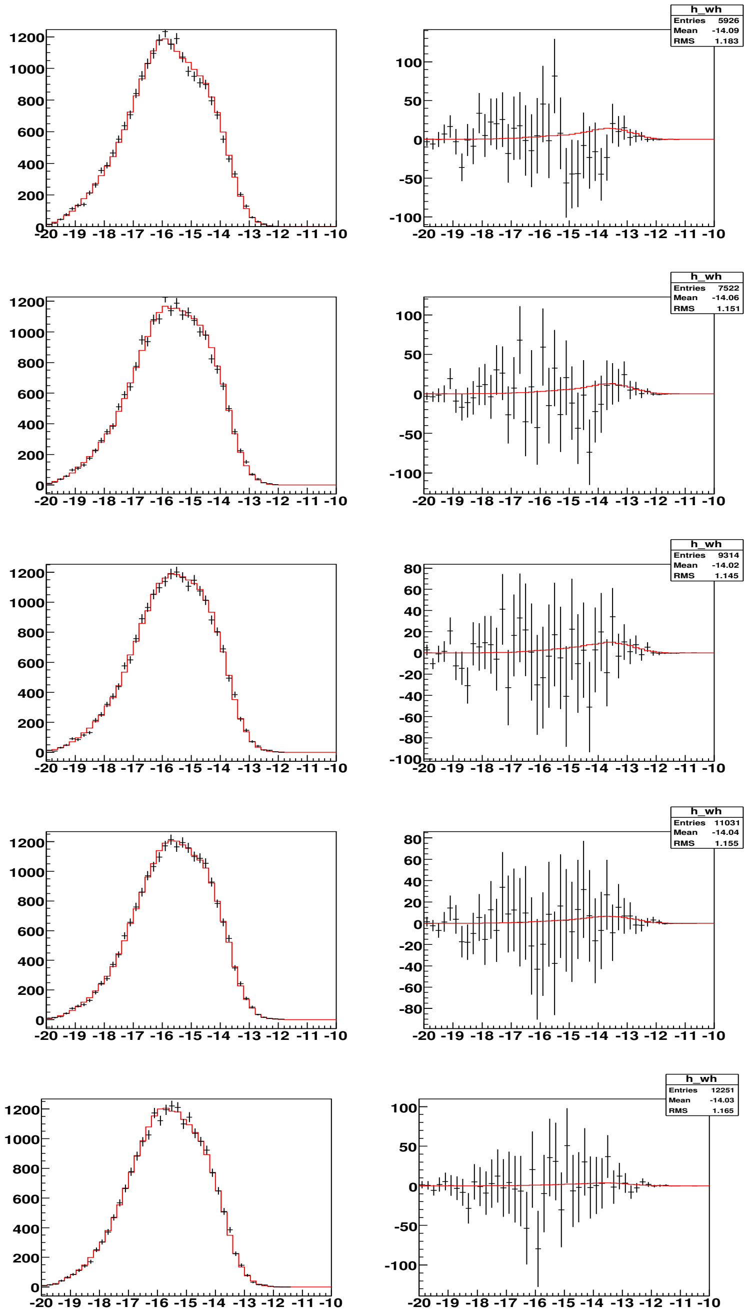

Figure 18: *Comparison of data and background prediction for (from top to bottom) Higgs masses of 100 through 140 GeV. Residuals are shown on the right; the line in the right plots shows the shape of the SM VH signal, multiplied by a factor 25.*

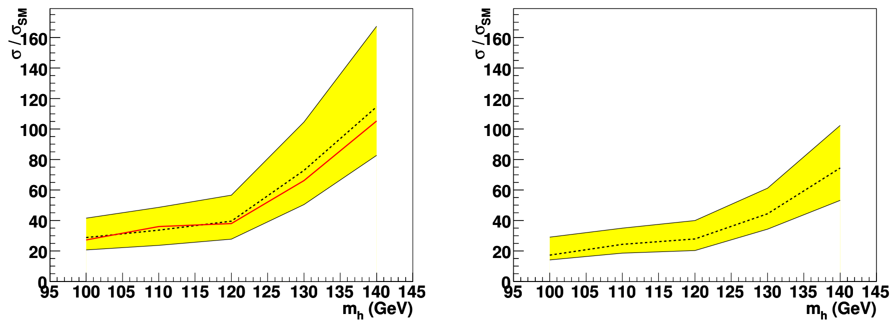

Figure 19: **left:** *Observed limit on the VH cross section as a function of Higgs mass.* **right:** *Expected limit when ignoring all systematic uncertainties.*

References

[1] The CDF Collaboration D. Acosta et al. Search for higgs bosons decaying into bb-bar and produced in association with a vector boson in p anti-p collisions at s**(1/2) = 1.8 tev. *Phys. Rev. Lett. 95*, 95:051801, 2005.

[2] Song-Ming Wang, Rong-Shyang Lu, Ankush Mitra, Aart Heijboer, Daniel Whiteson, Joe Kroll. A search for the higgs boson in the all hadronic channel using a boosted descision tree. CDF-9288.

[3] Song-Ming Wang, Rong-Shyang Lu, Ankush Mitra, Aart Heijboer, Daniel Whiteson, Joe Kroll. Study of the top multijet trigger efficiency for the search of the standard model higgs boson. CDF-9252.

[4] http://www.physics.ucdavis.edu/ conway/research/higgs/gg-vh-1960.txt.

[5] Michelangelo L. Mangano, Mauro Moretti, Fulvio Piccinini, Roberto Pittau, and Antonio D. Polosa. ALPGEN, a generator for hard multiparton processes in hadronic collisions. *JHEP*, 07:001, 2003.

[6] Fabio Maltoni and Tim Stelzer. MadEvent: Automatic event generation with MadGraph. *JHEP*, 02:027, 2003.

[7] Alexander L. Read. Modified frequentist analysis of search results (The CL(s) method). Prepared for Workshop on Confidence Limits, Geneva, Switzerland, 17-18 Jan 2000.

[8] M. Cacciari, S. Frixione, M. L. Mangano, P. Nason, and G. Ridolfi. The t anti-t cross-section at 1.8-TeV and 1.96-TeV: A study of the systematics due to parton densities and scale dependence. *JHEP*, 04:068, 2004.