1 Background Subtraction

Understanding the background to the $Z \to b\bar{b}$ signal in the double b-tagged data sample is crucial to this analysis. The background is composed nearly entirely of heavy-flavor di-jet production, and mis-tagged gluon/light-quark jet production, all of which is not possible to simulate accurately using current techniques. Thus the background will be derived from data, using either single-tagged and/or un-tagged events. Once the background shape is well understood, one can then subtract it from the data with confidence.

There are few kinematic handles with which to discriminate between $Z \to b\bar{b}$ signal and the QCD $b\bar{b}$ background. The two most powerful variables are the number of jets in the event, n_{jet} - which gives a measure of the amount of radiation from the b-jets - and the angular separation of the two leading b-jets, $\Delta\phi$. Intuitively, for a given jet p_T, one would expect the two leading jets to be more back-to-back in $Z \to b\bar{b}$ events than in QCD background, due to the presence of more gluon radiation in the QCD processes.

1.1 Invariant Mass Based TRF Method for Background Estimation (a.k.a. - The CDF Method)

One can attempt to estimate the background, therefore, by defining two regions in $n_{jet} - \Delta\phi$ space (DIAG HERE):

1. *"In Zone"* A region rich in signal. Events falling into this box pass our loose kinematic event selection (XXX), plus the additional requirements of $n_{jet} = 2$ and $\Delta\phi > 3.0$ radians. Requiring only two jets in the event restricts the amount of radiation between the jets while a tight $\Delta\phi$ cut ensures a more back-to-back topology.

2. *"Out of Zone"* A background-enriched region. Events pass only the loose event cuts. They fail the above 'in zone' n_{jet} and $\Delta\phi$ criteria.

The data is further divided into double-b-tagged and single-b-tagged samples. There are therefore four types of events:

- In Zone, double-tagged: N_{in}^{++}
- In Zone, single-tagged: N_{in}^{+}
- Out of Zone, double-tagged: N_{out}^{+}
- Out of Zone, single-tagged: N_{out}^{+}

In the signal-enhanced region, N_{in}^{++} is an admixture of genuine signal events and background events that have 'leaked' into this zone. The method initially adopted closely followed CDF's approach in Run I, using event counting in the invariant mass plane of the dijet system. The amount of background contamination in-zone is determined by constructing a tag-rate function (TRF) outside the zone, and then extrapolating this TRF into the zone. Here the TRF is defined as the ratio between N^{++}_{out} and N^{+}_{out}. The expected background in-zone is then evaluated as follows:

$$N^{++}_{exp-bkg,IN} = N^{+}_{obs,IN} \text{ x } \frac{N^{++}_{obs,IN}}{N^{+}_{obs,OUT}} \tag{1}$$

$N^{++}_{exp-bkg,IN}$ is then subtracted from the observed mass distribution after event selection. After subtraction, provided the number of events occupying the side-band regions is consistent with zero, an excess around the mass of the Z constitutes evidence for a signal.

The TRF is constructed in the dijet-mass plane and the estimated background is subtracted bin-by-bin from the invariant mass spectrum. Underpinning this method is the assumption that it is valid to extrapolate a TRF measured *outside* the signal zone *into* the signal zone. This in turn assumes that the TRF is independent of the choice of signal box, in particular of the choice of $\Delta\phi$ between the two leading jets in the event, and is discussed further in section 1.2.

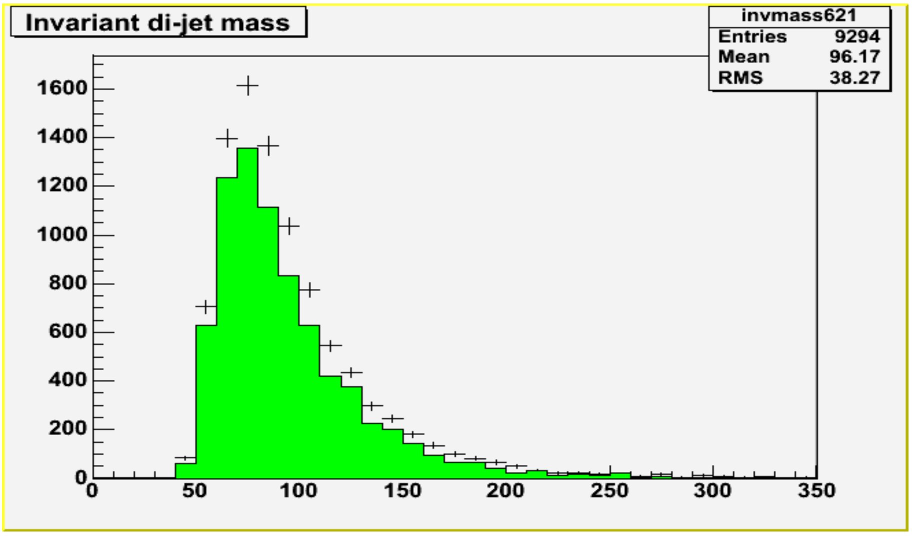

Figure 1: Invariant dijet mass distribution in the v12 dataset *before* background subtraction. Green shaded histogram: Expected background, estimated using an invariant-mass based TRF. Black points: Total dijet mass distribution observed in v12 data.

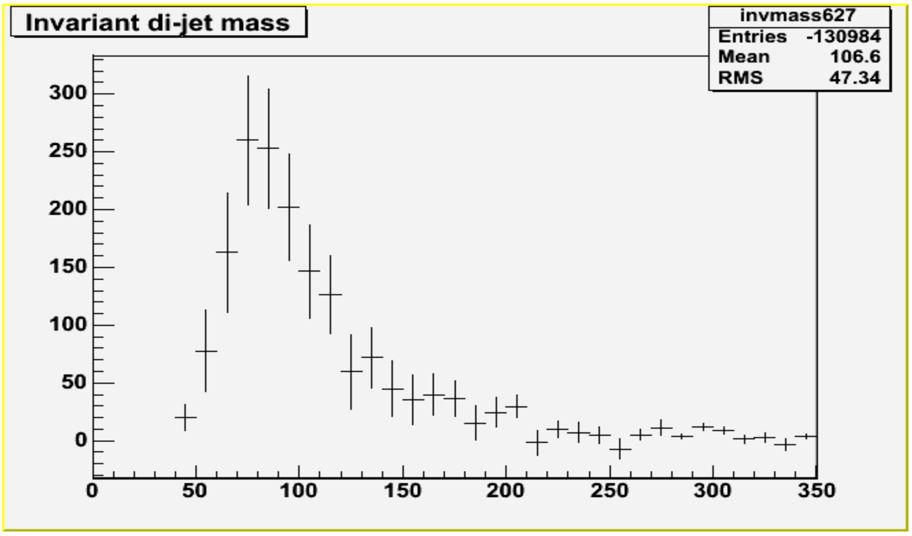

Figure 2: Invariant dijet mass spectrum from v12 data *after* background subtraction. The background is estimated using a tag-rate function calculated in the invariant-mass plane. An excess of events around the Z-peak is observed.

Figure 1 shows the observed events in black, overlaid with the expected background in green. There is an excess around the Z-peak, and the magnitude of this excess after background subtraction is shown in Figure 2. However, in Figure 1 it can also be seen that the background shape is not accurately modeled in the high-mass region. Outside the Z-peak region the data points lie consistently above the background estimate, which implies that the amount of background is being underestimated (both inside and outside the signal zone). The excess of events observed in Figure 2 after background subtraction is likely to be an overestimate.

1.2 Correction for the $\Delta\phi$ dependence of the TRF

Up to this point, it has been assumed that the ratio of double- to single-tags outside the signal zone is independent of where the signal box is placed, *id est* is independent of the particular $\Delta\phi$ value chosen to differentiate signal from background. Figure 3 suggests that this assumption is not valid, and that the TRF varies, depending on where in $\Delta\phi$ space it is measured.

In Figure 3, the red points show the variation of the TRF with dijet mass for signal-enriched events. The black points correspond to the TRF of events in a signal-depleted zone, where the jets are less back-to-back. As expected, the TRF of signal-like events peaks around the Z mass, while the

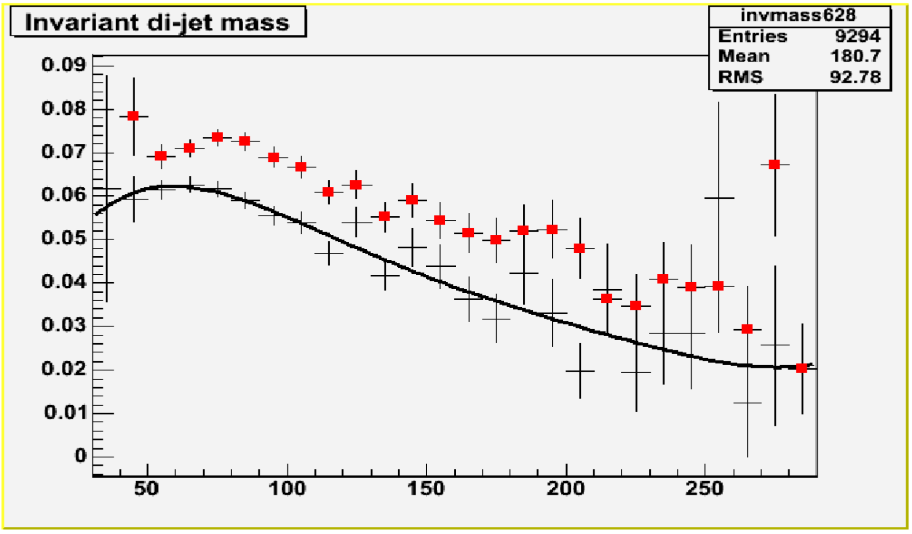

Figure 3: TRF as a function of invariant mass, for $\Delta\phi > 3.0$ radians (red points) and $2.8 < \Delta\phi < 3.0$ radians (black points). The distributions are derived from v12 data.

TRF of background-like events peaks at a lower mass. Away from the Z mass one would expect the two TRFs to converge, if the TRF is indeed independent of $\Delta\phi$. Such convergence is, however, not seen, implying that the ratio of double- to single-tags depends on the $\Delta\phi$ cut imposed.

Having demonstrated that a $\Delta\phi$ dependence exists, it is natural to evaluate and apply a correction for this dependence. We derive this by fitting a plot of TRF against $\Delta\phi$ from data, as shown in Figure 4. A linear dependence is seen in both Monte Carlo and data. As $\Delta\phi$ increases (*id est* the jets become more back-to-back and the event becomes more signal-like and less background-like), the ratio of double- to single-tags increases, as expected. A correction of 15% is observed for $2.8 < \Delta\phi < 3.0$ radians and this is used to scale up the estimated background.

This treatment assumes the signal is *only* found inside the signal zone, which means the signal estimate is conservative. In reality, some $Z \to b\bar{b}$ will be present outside the zone as the boundaries defining the signal box are blurred.

1.3 The Jet-Based TRF Method with $\Delta\phi$ Correction

We have seen that an invariant-mass based constant TRF does not model the background shape correctly at higher masses. In addition to correcting for the $\Delta\phi$ dependence of the TRF, we can improve the way in which the

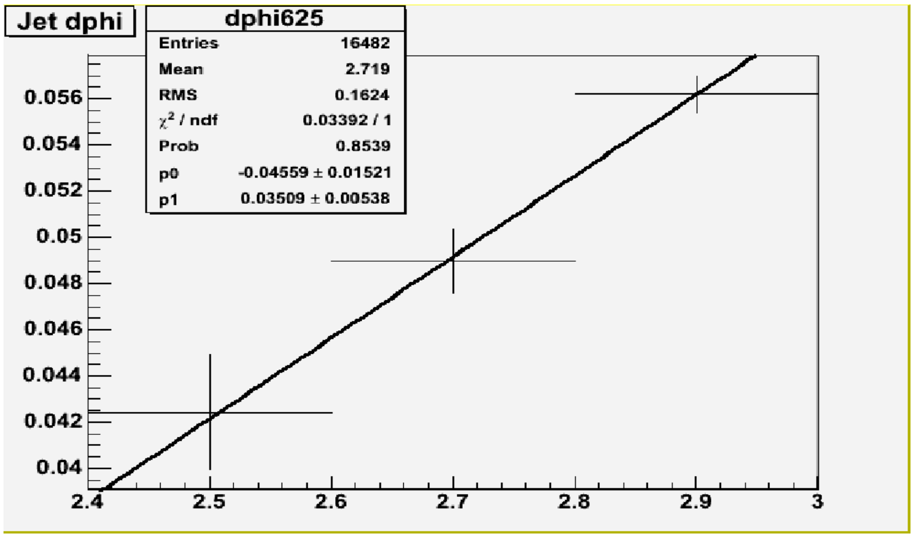

Figure 4: The variation of tag-rate function with $\Delta\phi$ between the two leading b-jets in each event, for v12 data. A linear dependence on $\Delta\phi$ is observed and the corresponding correction is derived.

TRF is calculated. Instead of an invariant-mass based TRF, a jet-based TRF can be constructed. This is a similar approach to that adopted in the DØ Run II $hb\bar{b}$ analysis [REF].

Consider events in which the first-leading jet is single-SVT tagged. For these events, the second leading jet is categorized according to its location in one of three different eta regions of the detector: XXX. For each η region, the TRF is then calculated as a function of the E_T of the second leading jet. This generates a TRF *per jet*. It is still based on events outside the signal zone.

Each event is then weighted accordingly. This is likely to be a more accurate method as it provides a finer resolution to the correction. The 15% correction for the $\Delta\phi$ dependence of the TRF is also incorporated. (How is it incorporated???)

The background subtraction method - a jet-based TRF with correction for $\Delta\phi$ dependence - is now applied to the v12 dataset. The absolute prediction for the amount of background in the signal zone is shown in the green shaded histogram of Figure 5. The black histogrammed points describe the actual distribution of events observed in data.

Insert MC peak here.

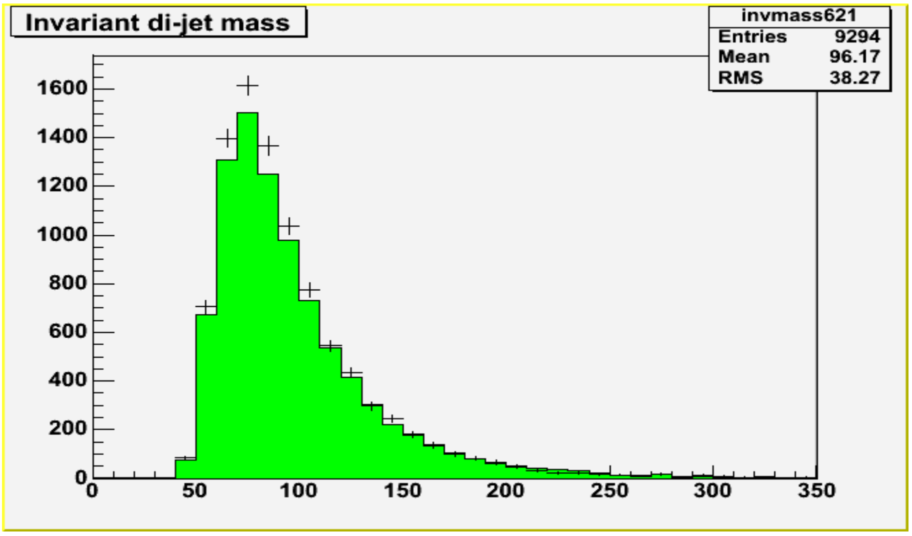

Figure 5: Invariant dijet mass in the v12 dataset *before* background subtraction. Green shaded histogram: Expected background, estimated using a jet-based TRF and scaling for the TRF dependence on $\Delta\phi$. Black points: total dijet mass distribution observed in data.

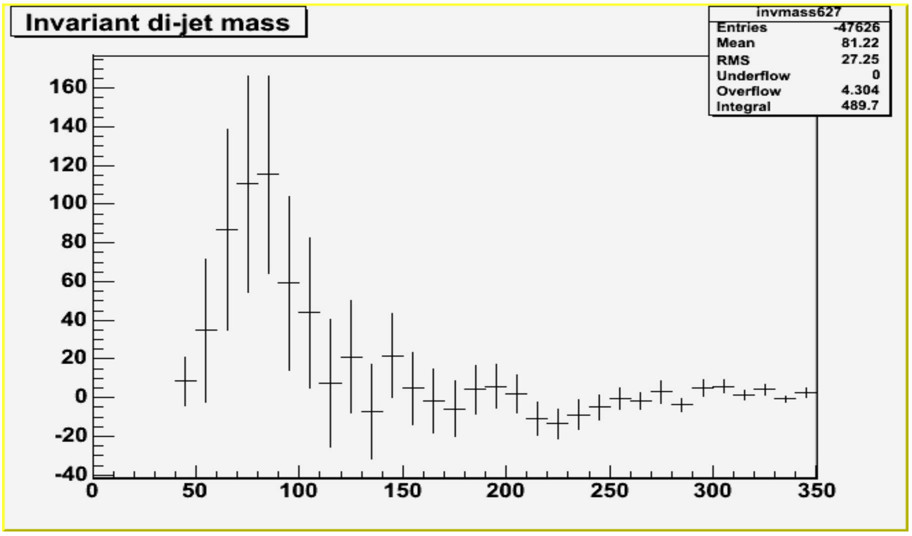

Figure 6: Invariant dijet mass in the v12 dataset *after* background subtraction. An excess of 490 $\pm$ 22 events around the Z mass is observed.

After background subtraction, an excess of 490 $\pm$ 22 events are observed around the Z-peak, which corresponds to a significance of 5. (We need to fix this error on the number of events... 22 is too low, and how is that 5 sigma?) The peak of the distribution lies at approximately 80 GeV/c^2, in both Monte Carlo and data, after full jet energy scale corrections have been applied. This suggests that the b-jet energy scale needs to be improved in order to fully reconstruct the mass of the Z boson.

1.4 The Full TRF Method with 0→1 and Z Corrections

Another method of background estimation will now be explained, which differs in the following ways from the previous method (the jet-based TRF method with $\Delta\phi$ correction).

- The method does not assume that the signal / background is dependent on $\Delta\phi$ or N_{jet}. The jet-based TRF is measured in, and applied to, the same data sample. (In the jet-based TRF method with $\Delta\phi$ correction, described above, different $\Delta\phi$ criteria were applied to the sample in which the TRF is measured than the sample to which the TRF is applied.)
- Because there is only one data sample used, no $\Delta\phi$ correction is applied to the TRF. This eliminates the systematic error which would be associated with that correction.
- The use of a single data sample allows the method to use larger statistics and include the signal events which would have been classified in the background-rich data sample by the previous method.

In addition, two corrections are applied to the background estimation:

- Because the TRF is applied to the single-tagged data sample, which is less rich in heavy-flavor jets than the data sample to which it is compared (the double b-tagged data), any differences in either the jet-energy-scale of heavy-flavor jets as opposed to gluon and light-quark jets or their natural invariant mass spectrum will result in a shift of the invariant mass distribution expected for the background, as compared to the data. The shift is observed, and measured, using *un-tagged* data sample. A TRF is derived for the un-tagged data and applied to the un-tagged data, and then compared to the actual single-tagged data. The shift which is derived is called the "0→1" correction, which is then subtracted from the expected background in the double-tagged data sample. This correction relies on the fact that each successive b-tag that is required increases the fraction of heavy-flavor by the same amount. This is verified by comparing the data with MC $b\bar{b}$ events.
- The effects of $Z \rightarrow b\bar{b}$ events which are present in both the un-tagged and single-tagged data samples from which the 0→1 correction and the TRF are derived, respectively, are corrected for. A signal peak is first measured in double-tagged data, using this background subtraction

method, including the 0→1 correction. The number of events which would exist from this signal peak in the un-tagged and single-tagged data samples is then extrapolated using a $Z \rightarrow b\bar{b}$ MC sample with no, one, and two b-tags required. The signal peak, *measured in data*, is then scaled by these factors and subtracted from the un-tagged and single-tagged samples. Then the TRF and the 0→1 correction are re-derived, the expected background in the double-tagged data is re-calculated, and a new signal peak is observed. This correction process is then repeated, using the new signal peak to estimate the $Z \rightarrow b\bar{b}$ events in the un-tagged and single-tagged data samples, a total of three times, after which the correction and signal peak are stable.

1.4.1 Event Samples and Selection

The data sample used for testing the full-TRF method is also the "BID" skim of the "Pass2" dataset, which has the following properties:

- A "loose" muon in each event, with p_T >4 GeV, matched to a JCCB (0.5 RunII cone) jet within ΔR <0.7. This requirement enhances the fraction of heavy-flavor events due to the decays of $b \rightarrow \mu$ and $b \rightarrow c \rightarrow \mu$. (This was the best skim available at the time of this analysis. For the future, a new skim has been proposed ("2JET") which requires exactly two JCCB jets with $p_T > 15$ GeV and does not demand any muons. Initial studies indicate that the total efficiency for signal events will increase by 50–100% using the new skim.)
- The skim contains about 90M events and has an integrated luminosity of about $300 pb^{-1}$. (The integrated luminosity can only be roughly approximated, since the data are not selected by any given set of triggers, and the prescales and overlaps are largely unknown. The integrated luminosity quoted is estimated by comparing the invariant mass spectrum of di-jet events in the full data sample to that obtained from a smaller data sample selected by requiring a trigger with known luminosity.)

In addition, these "Pass2" MC samples are used:

- About 200K events of $Z \rightarrow b\bar{b}$, generated with Pythia.
- About 200K events of direct $b\bar{b}$, generated with Pythia, with a cut of $p_T > 40$ GeV applied at the parton-level. (While this p_T cut is certainly high compared to the Z peak, the sample is still useful for comparing the invariant mass spectrum in data to that in MC above ≈80 GeV.)

The data skim and MC sample were processed through the higgs_multijet rootuple maker. Events were then corrected and selected as follows:

- A good luminosity block was required for data quality
- A good primary vertex was required, with >=4 tracks attached, and $|z|$ <35 cm (the fiducial b-tagging region)
- Jet quality cuts derived by the Jet ID group for Pass2 data and MC were applied to all jets
- Exactly two jets were required
- Jet Energy Scale corrections, version 5.3, were applied to all jets, including the corrections for muons in jets
- The two jets were required to be back-to-back in ϕ, with $\Delta\phi$ >2.9 radians. (This cut could probably be lowered to ≈2.5 radians. Beyond that, processes involving $g \rightarrow b\bar{b}$ and $g \rightarrow c\bar{c}$ splitting are difficult to subtract from the double b-tagged data.)
- For b-tagging, the SVT "loose" criteria were applied, using the standard p14 definitions

1.4.2 The $0 \rightarrow 1$ tag Shift

As described above, the shift in the invariant mass spectrum of the di-jet system caused by applying a single b-tag is first measured. Later this will be subtracted from the estimated background of the double b-tagged data. A TRF is derived on the un-tagged sample and re-applied to the same un-tagged data sample, to predict the background to the single-tagged data sample, as shown in Figure 7. A comparison is also shown in this figure to the $b\bar{b}$ MC di-jet invariant mass spectrum (which was normalized using double-tagged data). The $b\bar{b}$ events make up about 10% of the single-tagged data sample, as opposed to about 2% in the un-tagged data sample. The small contribution of $Z \rightarrow b\bar{b}$ expected in the single-tagged data is also shown. Most of the events in the single-tagged data sample contain only gluon/light-quark jets.

The difference between the single-tagged data and the expected background from the TRF is shown in Figure 8. A comparison of the background-subtracted single-tagged data to the MC $Z \rightarrow b\bar{b}$ invariant mass distribution shows that it is not the result of a signal peak in the single-tagged data, but rather an overall shift in the invariant mass distribution. As mentioned above, this shift is understood to come from either a difference between the gluon/light-quark and the b-quark jet energy scale factors or a difference in the ratio of true cross-section of heavy-flavor jets to light-jets as a function of di-jet invariant mass.

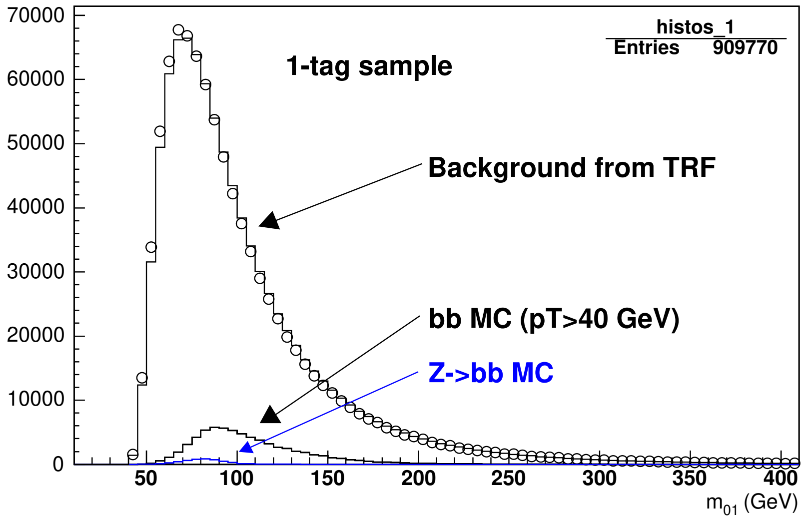

Figure 7: Comparison between the single-tagged data and the background expected using the TRF method. Comparisons are also shown to the $b\bar{b}$ MC di-jet invariant mass spectrum (which was normalized using double-tagged data) and to the $Z \to b\bar{b}$ MC di-jet invariant mass spectrum, to give a feel for the composition of the sample. The rest of the events are thought to be gluon/light-quark jet events.

1.4.3 The Double-tag TRF

To estimate the double b-tagged background, a set of TRFs are derived as a function of corrected jet p_T for 3 η-bins from the single-tagged data and reapplied to the same data. The TRFs are shown in Figure 9. The resulting estimate of the double-tagged background is compared to double-tagged data and shown in Figure 10.

1.4.4 The 0- and 1-tag Z Peak Correction

As mentioned above, the un-tagged and single-tagged data samples contain signal events, which will now be corrected for. The signal peak observed in double-tagged data, after subtracting the estimated background using the TRF method, is scaled by a factor of 6.5, which is the ratio of single-tagged to double-tagged events in the $Z \to b\bar{b}$ MC sample. It is important to note that the MC is only used for an overall normalization (the b-tagging efficiency), and does not affect the shape of the signal peak. This scaled signal peak is shown in Figure 11.

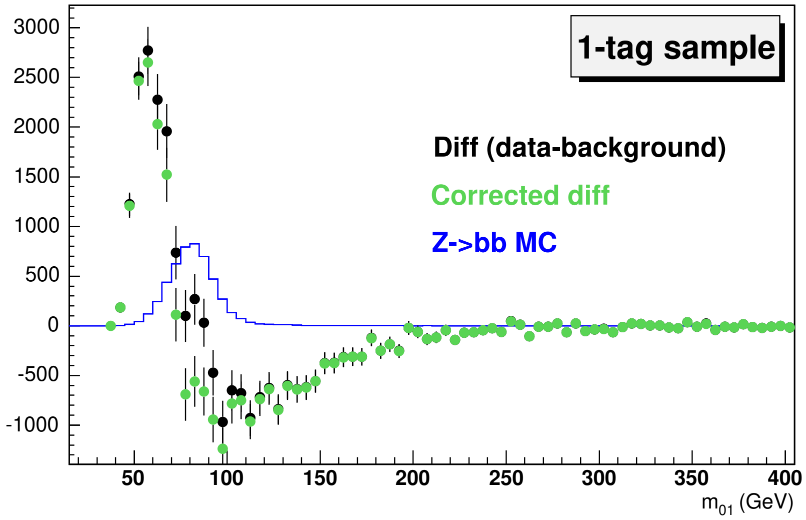

Figure 8: Difference between the single-tagged data and the background expected using the TRF method, a measure of the 0→1 tag shift, which will be subtracted from the expected double b-tagged data (after proper normalization). The difference is also shown after correcting for the expected $Z \to b\bar{b}$ events in the un-tagged and single-tagged data samples, using the methods described below in Section 1.4.4.

The expected fraction of $Z \to b\bar{b}$ events in each invariant mass bin is then subtracted from the signal-tagged data (the events are weighted by 1-f, where f is the fraction of $Z \to b\bar{b}$ expected). Then the TRFs are re-derived and re-applied to this corrected, weighted, single-tagged data. The 0→1 tag correction is also re-derived, using the same principles to estimate the fraction of $Z \to b\bar{b}$ in the un-tagged data sample. The effect from $Z \to b\bar{b}$ events in the un-tagged sample is relatively small, as expected, so the modification to the 0→1 tag correction is slight, as seen in Figure 8.

Since the size and shape of the original Z peak in data changes slightly after the corrections to the un-tagged and single-tagged data, the corrections themselves can be re-derived using the new Z peak observed. This cyclic process could be continued ad-infinitum, but in practice the solution is stable after only a few iterations, which results in a final Z peak and set of corrections.

1.4.5 Results and Conclusions of the Full-TRF Method

After Z peak corrections to the un-tagged and single-tagged data samples, the comparison of the background-subtracted double-tagged data to the

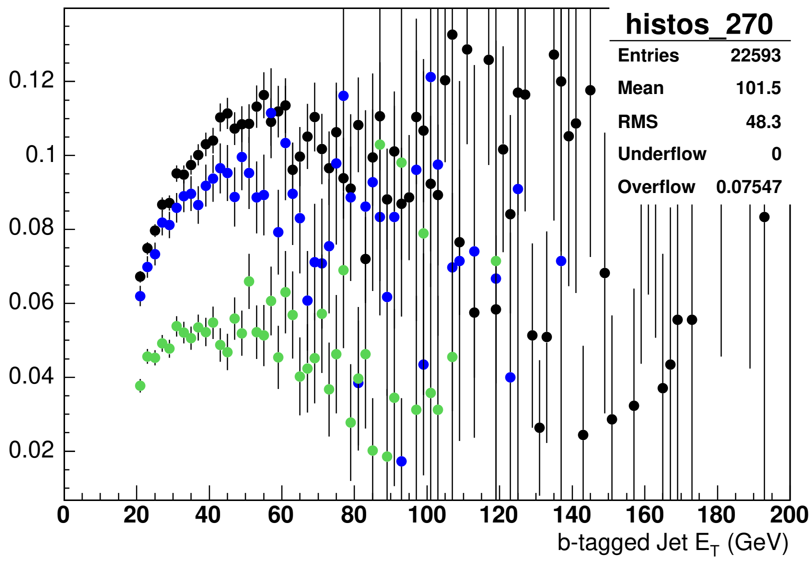

Figure 9: The TRF's derived on the single-tagged data sample, which are used to estimate the double-tagged background. Each TRF is a function of corrected jet p_T in one of 3 η-bins: $\eta < 1.1$, $1.1 < \eta < 1.5$, and $1.5 < \eta < 2.5$.

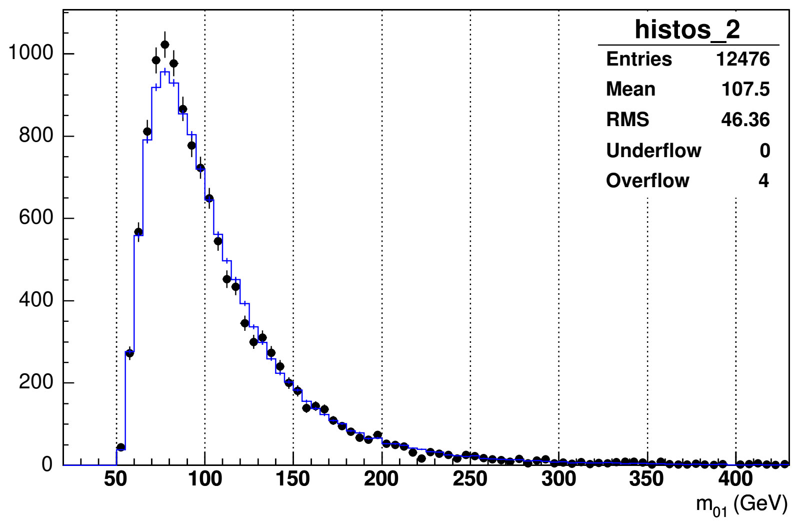

Figure 10: The comparison between double b-tagged data (points) and the expected background, from the TRF method using single-tagged data, before either of the background corrections ($0 \rightarrow 1$ or Z peak).

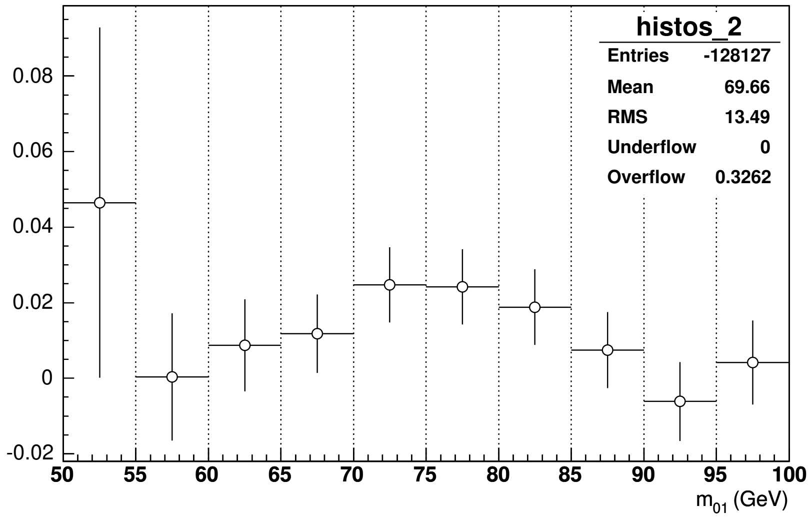

Figure 11: The estimated fraction of $Z \to b\bar{b}$ events in the single-tagged data sample, as estimated from the peak observed in the double-tagged data sample after subtracting the background estimated via the TRF method. This fraction of events will be subtracted from the single-tagged data, and then the final TRFs are derived and re-applied to this corrected single-tagged data set.

observed 0→1 tag shift is shown in Figure 12. (The slightly different shape of the 0→1 tag shift without Z peak corrections to the un-tagged and single-tagged data samples is also shown, to give a feel for the size of the effect.)

The final $Z \to b\bar{b}$ peak derived from data, after all corrections, is considered to be the difference between the background-subtracted double-tagged data and the corrected 0→1 tag shift. This invariant mass distribution difference is shown in Figure 13, and compared to the shape of the $Z \to b\bar{b}$ distribution in MC. (The slightly different shape of the difference between the background-subtracted double-tagged data and the 0→1 tag shift without Z peak corrections to the un-tagged and single-tagged data samples is also shown, to give a feel for the size of the effect.)

The final Z peak in data (with all corrections included) is fit to a Gaussian distribution, which describes the shape of the $Z \to b\bar{b}$ peak well in MC. Both the mean (79±3 GeV) and width (9±3 GeV) of the distribution observed for data are comparable to those derived from $Z \to b\bar{b}$ MC: 81 GeV and 11 GeV, respectively.

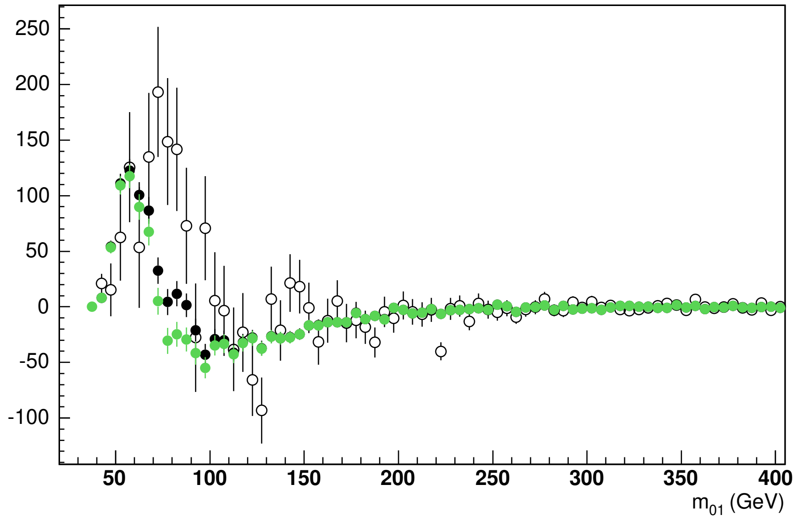

Figure 12: Comparison of the background-subtracted double-tagged data to the observed 0→1 tag shift. The slightly different shape of the 0→1 tag shift without Z peak corrections to the un-tagged and single-tagged data samples is also shown, to give a feel for the size of the effect.

810±230 $Z \rightarrow b\bar{b}$ events are observed, where the error is taken from the uncertainty in the height of the Gaussian fit. Recall that the data set used to test this background estimation method is able to use a larger $\Delta\phi$ range (2.9 as opposed to 3.0), spans the pre-v12 trigger-list data in the Pass2 skim, and corrects for the signal events in the data from which the TRFs were derived. (570±160 events are observed if the signal events in the data from which the TRFs were derived are not corrected for.) The total number of events expected from $Z \rightarrow b\bar{b}$ MC can not be estimated very accurately, because the data were not selected by a given trigger. However, for an assumed luminosity of $300pb^{-}1$, the size of the $Z \rightarrow b\bar{b}$ peak observed in data would correspond to an overall trigger/skimming efficiency of about 15%. This efficiency is reasonable since the trigger/skimming require a muon, and the signal has two b-jets, each with a 10% branching ratio to muons.

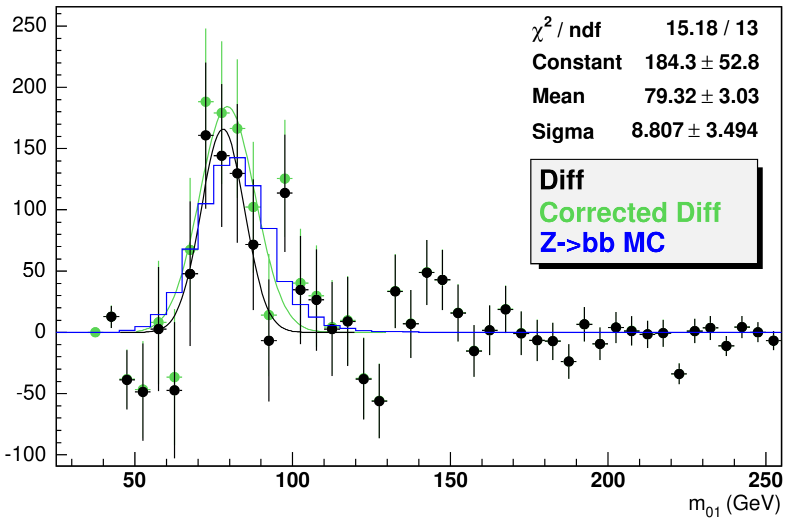

Figure 13: The final $Z \to b\bar{b}$ peak derived from data, after all corrections, compared to the shape of the $Z \to b\bar{b}$ distribution in MC. (The slightly different shape of the Z peak in data without the Z peak corrections to the un-tagged and single-tagged data samples is also shown, to give a feel for the size of the effect.)