CDF/PHYS/BOTTOM/CDFR/7318
October 28, 2004
Version 1.3

Monte Carlo study of lepton SVT sample

G. J. Barker, M. Feindt, U. Kerzel, C. Lecci
Univertsität Karlsruhe

G. Bauer, G. Gomez-Ceballos, I. Kravchenko, N. Leonardo,
S.Menzemer, Ch. Paus, A. Rakitin
Massachussetts Institute of Technology

J. Piedra, A. Ruiz, I. Vila
Instituto de Fisica de Cantabria

I. Furic
University of Chicago

Abstract

In the following note we present a study of event properties for a lepton SVT trigger sample. Our purpose is to understand the sample in aspects that are relevant for B flavor tagging and thus for measurements of CP violation, B^0 and B_s mixing. In particular we focus on understanding of the hemisphere away from the trigger lepton that is relevant for Opposite Side Taggers. The trigger side is also investigated.

We show the evidence that the Pythia Monte Carlo with only $b\bar{b}$ flavor creation process does not provide an adequate description of data. We propose a Monte Carlo sample with all production mechanisms and show the improvement in terms of data description.

With this sample we proceed investigating questions relative to Opposite Side Tagging, such as a choice of jet clustering algorithms.

Finally, we compute the purity of the selection algorithms used for the Jet Charge Tagger described in CDF note #7131.

1 Introduction

In this note we study the features of CDF events triggered by the ℓ+SVT trigger. We investigate the definition of event "sides", their characteristics and their correlations.
In each event the sum of the momenta of the trigger tracks identifies the *signal B candidate*. The *same side* is defined as the cone angle of size 0.7 around the signal B candidate direction. The event *opposite side* is then the detector volume not included in the same side.
The most important element of our study is the Monte Carlo sample. It helps us to evaluate distributions and to compute efficiency and purity of jet clustering and jet selection algorithms.
The Monte Carlo generated with Pythia [1] traditionally used in the CDF B group contains only $b\bar{b}$ pair production. From our study it emerges that this Monte Carlo does not describe the features of the opposite side accurately. The reason is that the Pythia $b\bar{b}$ pair production does not include other important $b\bar{b}$ creation processes.
We propose here an alternative sample composition, including additional $b\bar{b}$ creation processes, and compare it to the traditional sample. We show that the alternative sample gives a better description of the data and we use it to study the characteristics of the opposite side .
The results found can help to find the best way to identify the decay products of the opposite side b-hadron with jet clustering algorithms. They also suggest how difficult the task of selecting a tagging jet in ℓ+SVT triggered events can be. The correlation we found between opposite side and the trigger lepton side can help in understanding the measured dilution of opposite side taggers.
We describe in section 2 our Monte Carlo samples. Section 2.1 explains the technical details of the samples generation, simulation and reconstruction. The differences at generator level between the samples are shown in section 2.2. A comparison of opposite side related quantities among Monte Carlo samples and data sample is given in section 2.3 and an example of the quality of the detector description given by the simulation is given in section 2.4.
Section 3 describes the features of the same side (3.1) and the opposite side (3.2) of Monte Carlo ℓ+SVT events. An evaluation of the clustering algorithms currently in use in the B group is presented in section 4. Section 5 shows the purity of the tagging jet selection algorithm used for the opposite side Jet Charge Tagger [2]. Finally, section 6 draws the conclusion.

2 Choice of Monte Carlo samples

The study presented in this note starts from the observation that a Monte Carlo simulation of $b\bar{b}$ events produced only via leading order processes does not reproduce $b\bar{b}$ events as they are seen in the CDF detector. Some studies [3] [4] show that, in order to predict the right $b\bar{b}$ production cross section at the Tevatron, two Next to leading order processes have to be added to a leading order Monte Carlo. The Feynman graphs associated to these production mechanisms are shown in Fig.1.
Flavor creation (FC) is the leading order process. It creates the heavy quark pair through one of the diagrams in Figs. 1a and 1b. The emission of a gluon by one of the heavy quarks (Fig. 1c) does not affect the cross section of the process [3].
The Next to leading order processes are *flavor excitation* (FE) and *gluon splitting* (GS).
Flavor excitation happens when a virtual heavy quark from the parton distribution of one incoming beam particle is put on mass shell by the momentum transferred through the interaction with a parton in another beam particle (Fig. 1d).
In gluon splitting processes no heavy quark takes part in the hard scattering. The $b\bar{b}$ pair is produced either by a final or initial state gluon (Figs. 1e and 1f respectively).

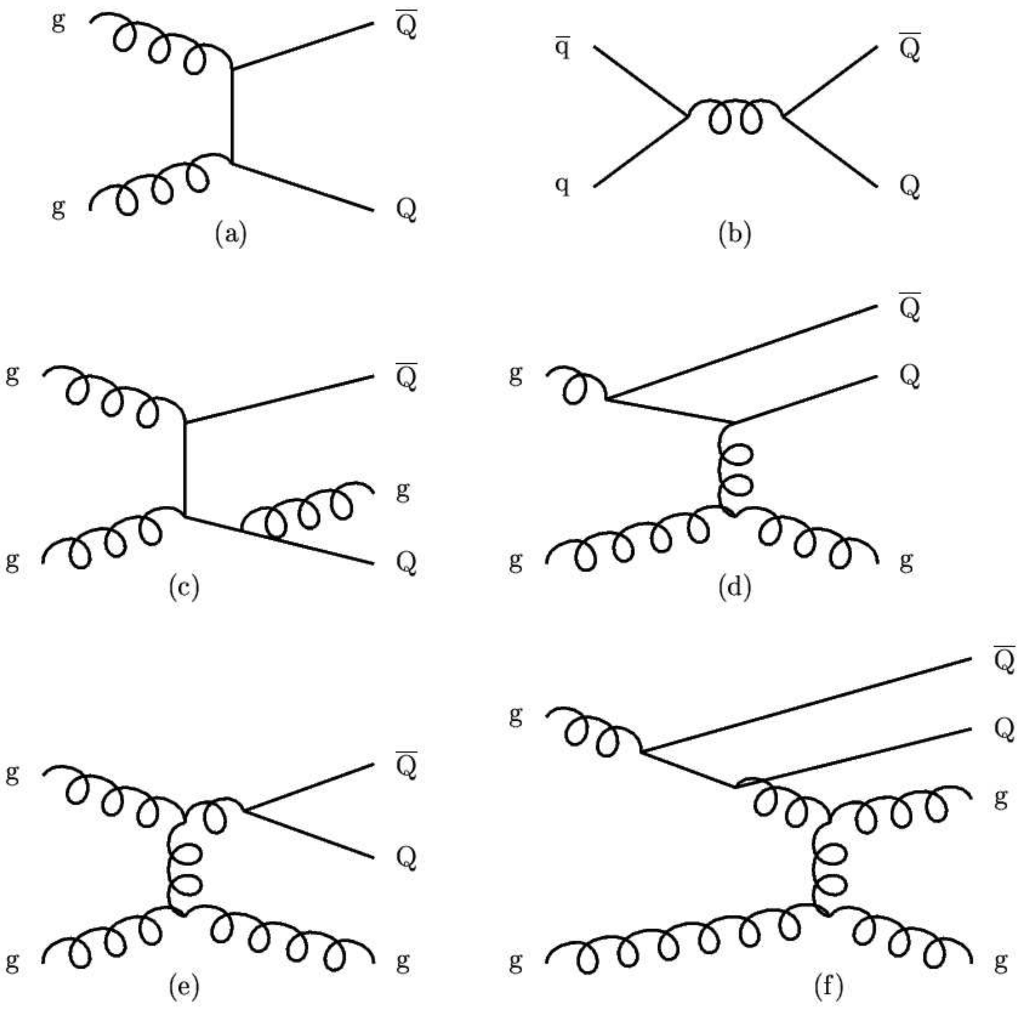

Figure 1: Feynman diagrams of the most important $b\bar{b}$ production processes. *(a)*, *(b)* and *(c)* show flavor creation processes. *(d)* is a flavor excitation diagram. *(e)* and *(f)* correspond to gluon splitting processes [3].

The three processes result in different correlations between the produced quarks [4].
The study [3] shows that at 2 TeV center of mass energy the $b\bar{b}$ production cross section by flavor excitation processes is comparable to the cross section of flavor creation. Gluon splitting plays a less important role but is, nevertheless, not negligible.
All three of these $b\bar{b}$ production processes have to be included in the generation of a Monte Carlo sample to obtain a reasonable description of the data.

2.1 All processes ℓ+SVT Monte Carlo sample description

A sample containing $b\bar{b}$ flavor creation only processes and a sample containing all $b\bar{b}$ generation processes was generated.
The Monte Carlo sample with flavor creation only has been generated, simulated and reconstructed at the GridKa computing facility. The sample consist of 140k $b\bar{b}$ events triggered by the lepton and displaced track trigger.
The sample has been generated with Pythia v6.2 with the option `msel set 5`, that is $b\bar{b}$ flavor creation, with non zero b quark mass in the leading order matrix elements.

The decay package used was QQ [5].
The version of CDF software used to produce this sample was 5.3.0. The `cdfSim` executable was a custom version based on 5.3.0 and including the latest version of ToF reconstruction available at the time of the generation (end of April 2004).
The set of tcl switches is the default available for 5.3.0 release. The flag `cdfSim_SI_Matching` has been set to 1 in the `cdfSim` steering file in order to keep all the data banks needed for track matching to Monte Carlo particles.
The events have been simulated in the run range 138815 - 156487, corresponding to the data taken up to the shutdown in January 2003 and the sample contains about 140k events.
The Monte Carlo sample containing all the $b\bar{b}$ production processes has been produced on the Fermilab CAF by using the `nbot90` sample [6]. The latter sample consists of 25M events containing only the four vector quantities of the generated particles. The events have been generated with Pythia with the option `msel set 1`, i.e. the production of all quark flavors with all processes. A filter has been applied to select only events containing at least a b or a $\bar{b}$ quark with transverse momentum $p_T > 4$ GeV and pseudorapidity $|\eta| < 1.5$.
The four vector events have been processed by EvtGen [7] and simulated. The ℓ+SVT trigger has been simulated and the events surviving after the filter have been reconstructed. Since the trigger filter drastically reduces the initial number of events, the `nbot90` sample has been re-decayed and simulated for several iterations. The probability that the same initial event passes the trigger twice is very small. The duplicated events, which amount to a 1% fraction of the total number of simulated events, have been tabulated so that it is possible to exclude them in the DHInput module.
The number of events in this sample is about 113k. The simulated run range is 138809 - 178785, which corresponds to the list of good runs taken up to Feb. 2004.
The version of CDF software used to produce this sample was 5.3.3. The `cdfSim` executable has been compiled in order to include the latest version of ToF reconstruction available (`toftag2`).
The default parameters of the 5.3.3 release have been used for simulation and reconstruction. Also for this sample the flag `cdfSim_SI_Matching` has been set to 1 in the `cdfSim` steering file.
The information about the particular hard scattering process that happened in each event is lost after the generation. It is not possible to state with certainty if in a given event the $b\bar{b}$ pair was produced by flavor creation, flavor excitation or gluon splitting.
All plots and numbers that will be discussed in the following sections were prepared removing the events in which the invariant mass of the lepton and SVT track is below 2 GeV or above 4 GeV [8]. The background subtraction was performed as in [9].

2.2 Correlation of b and $\bar{b}$ quarks

In each event of each Monte Carlo sample we reconstruct the signal B candidate fitting a vertex with the trigger lepton and displaced track. The procedure to reconstruct the B candidate is the same as in the Jet Charge analysis [9] .
In each simulated event we identify the heavy quarks from the Monte Carlo banks OBSP, OBSV and HEPG. The procedure is the following: first, the two produced b-hadrons are found in the Monte Carlo stable particles bank, OBSP; then the angle in space between each b-hadron momentum and the reconstructed signal B candidate momentum is computed. The b-hadron with the smallest angle is the *signal B*, the other is the *tagging B*.
For each of the b-hadrons, the corresponding b quark is found in the HEPG bank by going up in the generation tree. The quarks are accordingly called *signal quark* and *tagging quark*. The correlation between the signal and the tagging quarks is displayed in the plots in Fig.

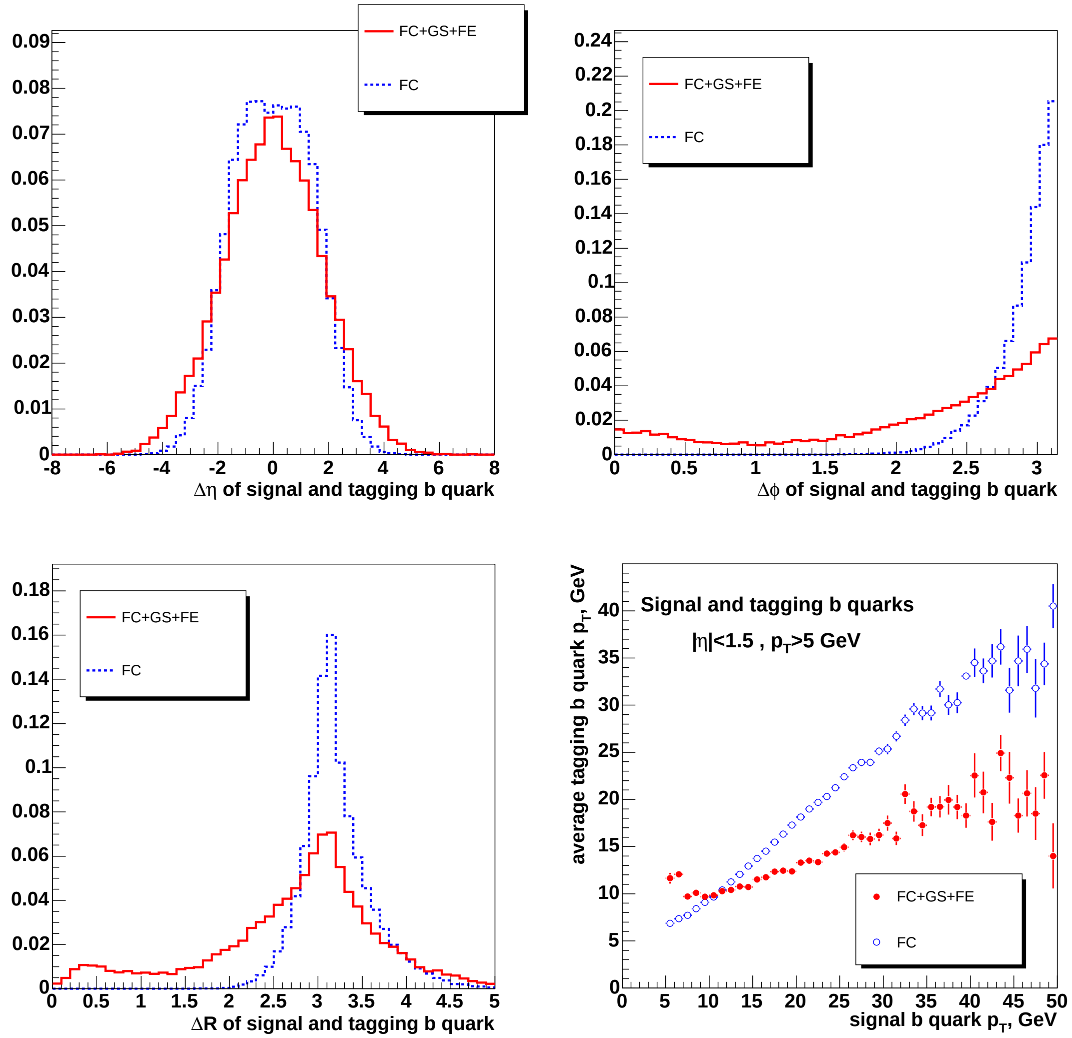

Figure 2: $b\bar{b}$ correlation plots. The signal b quark is required to be in the detector acceptance ($|\eta| < 1.5$, $p_T > 5$ GeV). *Top left:* pseudorapidity difference between the quarks. *Top right:* angular difference in the transverse plane. *Bottom left:* cone angle $\Delta R = \sqrt{\Delta\eta^2 + \Delta\Phi^2}$ between the quarks *Bottom right:* p_T correlation.

2. For each of these plots, the signal quark is in the detector acceptance ($|\eta| < 1.5$) and has a minimum pT of 5 GeV. No cuts are applied on the tagging quark for the plots on the difference of pseudorapidity ($\Delta\eta$, Fig. 2, top left), azimuthal angle ($\Delta\Phi$, Fig. 2, top right) and cone angle (ΔR, Fig. 2, bottom left). The η and p_T cuts are applied to both quarks in the p_T correlation plot (Fig. 2, bottom right). Each histogram is normalized to the same area. The same normalization has been used throughout the note.

In flavor creation processes the b quarks are produced preferentially back to back and the correlation between their momenta is strong. This does not hold for gluon splitting and flavor excitation. In the latter two cases the quark momenta are less correlated, as one can see in the p_T correlation plot.

In gluon splitting processes it can happen that the quarks fly very close in direction, thus making impossible the definition of same side and opposite side. In this case, it is difficult to properly flag one of them as signal and the other as the tagging quark and wrong assignments can happen. The p_T cut on the signal side biases the transverse momentum of the signal

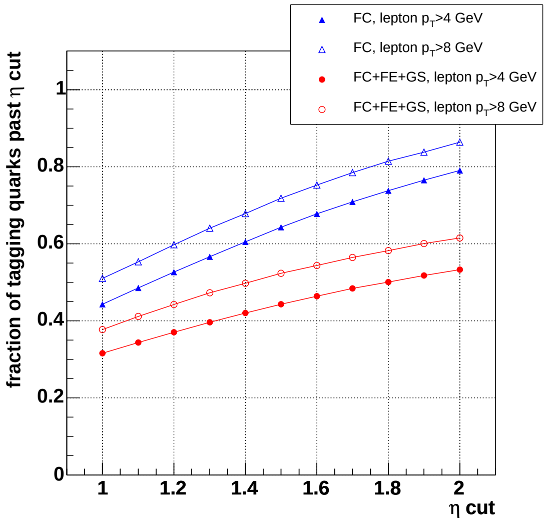

Figure 3: The signal b quark is required to have $p_T > 5$ GeV and $|\eta| < 1.5$. The plot shows the fraction of events in which the tagging b quark has $p_T > 5$ and $|\eta| < \eta_{cut}$. Different curves are drawn for the two Monte Carlo samples and for two different p_T cuts on the trigger lepton.

quark to higher values. In case of wrong assignments the bias is on the tagging side and this explains the rise at low momentum that one observes in the p_T correlation plot.
We consider only the events of the Monte Carlo samples in which the signal b quark satisfies the cuts $|\eta| < 1.5$ and $p_T > 5$. The fraction of these events for which the tagging quark quarks has $|\eta| < \eta_{cut}$ and $p_T > 5$ GeV is shown in Fig. 3 for different values of η_{cut}.
In general the flavor creation Monte Carlo (full triangles) overestimates the fraction by 15% to 30% with respect to the sample with all processes (full circles). For $\eta_{cut} = 1.5$ the fraction is 44% for the all processes sample and 64% for the flavor creation sample.
We apply an additional cut of 8 GeV on the trigger lepton p_T (open triangles and circles in Fig. 3). The fraction of events passing the η cut becomes about 10% larger in both samples. This increase could justify the fact that the Jet Charge Tagger achieves a better performance on a 8 GeV lepton sample than on a 4 GeV lepton sample [9].
If we require also that the tagging quark is outside a cone $\Delta R = 0.7$ around the signal quark direction (*isolation cone*) the curves relative to the all processes sample are shifted downward by 3%. The isolation cone cut does not affect the flavour creation sample curves.
The comparison between the two Monte Carlo samples at generator level points out some differences that play an important role when studying the opposite side of ℓ+SVT events.
According to the flavor creation Monte Carlo the heavy quarks are well separated in space and are both in the detector for most of the triggered events. Once the signal hadron direction is reconstructed in the transverse plane, the tagging hadron is likely to be found by looking in the opposite direction.
The all processes Monte Carlo offers a more complex view. The tagging quark is mostly out of the detector acceptance and sometimes very close in space to the signal quark. We can expect that the chance of defining correctly the opposite hemisphere and identifying the tagging hadron is consequently smaller in this sample.

2.3 Data/Monte Carlo comparison of opposite side quantities

In the following section we compare real data to Monte Carlo distributions of some opposite side related quantities. The data sample used is processed with version 5.1 of the reconstruction software and is a subsample of `jbel0c` and `jbmu0c`.

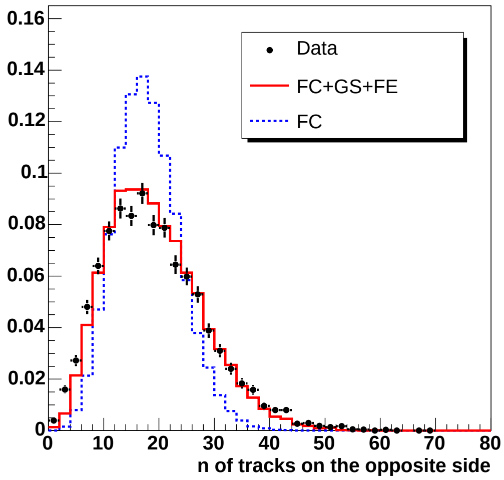

Figure 4: Number of tracks in the event outside a cone of $\Delta R = 0.7$ around the B candidate direction. The data points are compared to the distributions obtained for flavor creation only Monte Carlo and for all processes Monte Carlo.

The signal B candidate is reconstructed in data in the same way as in the simulated events (see 2.2).

The opposite side in each event is defined as the detector volume outside a cone of 0.7 around the reconstructed B candidate direction.

We apply the Cone Clustering algorithm [10] on the opposite side to reconstruct jets and select among them a tagging jet with the JetSelection algorithm [11]. The parameters used for Cone Clustering and JetSelection algorithm are the tuned values found in [9], in particular a cone size of 1.5 is used.

Every distribution shown in this section and in the following ones has been normalized to the same area.

The number of tracks reconstructed on the opposite side (Fig. 4) is better modeled in the Monte Carlo containing all processes than by the flavor creation only Monte Carlo. The average number of opposite side tracks is very similar in the two Monte Carlo samples, but the width of the distribution is larger for the all processes Monte Carlo sample. This suggests that one of the two next to leading order processes generates in average more particles than flavor creation and the other process generates fewer particles.

The number of reconstructed jets on the opposite side by the Cone Clustering algorithm (Fig. 5, top left) is in average smaller in the all processes Monte Carlo than in the flavor creation only Monte Carlo. The overall distribution for the first sample is closer to the data.

The distribution of the number of tracks in the tagging jet (Fig. 5, top right) for data is more compatible with the all processes Monte Carlo distribution than with the flavor creation only Monte Carlo. This is in agreement with the better description of track multiplicity on the opposite side given by the all processes Monte Carlo.

Finally, the distributions of the angle in the transverse plane between the tagging jet and the signal B candidate direction (Fig. 5, bottom left) and the cone angle ΔR between the same directions (Fig. 5, bottom right) clearly show that the all processes Monte Carlo reproduces the opposite side features of data events significantly better than the flavor creation only Monte Carlo.

In the following sections the flavor creation only sample will be abandoned and the data

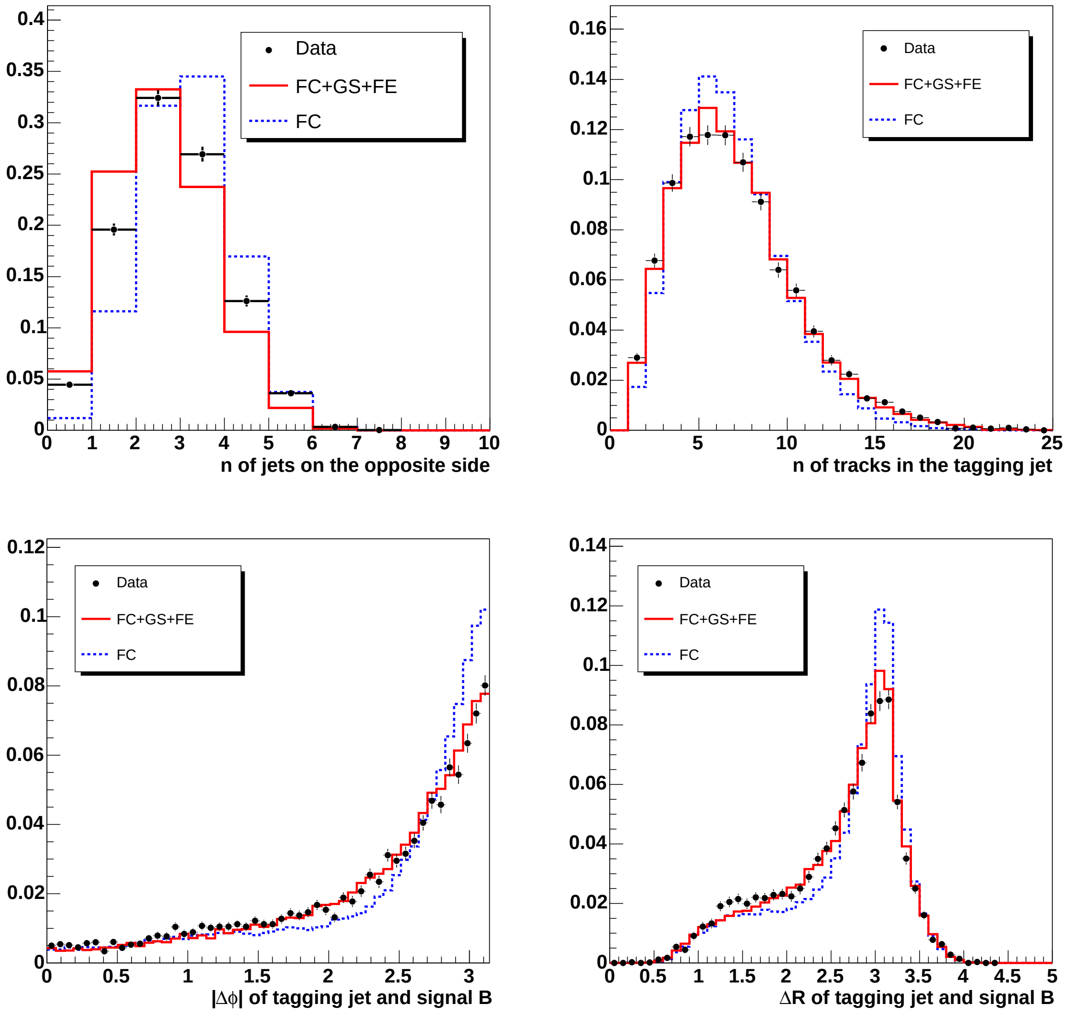

Figure 5: Comparison of distributions of opposite side variables for data, flavor creation Monte Carlo and all processes Monte Carlo. *Top left:* Number of reconstructed jets on the opposite side by the cone clustering algorithm optimized as in [9]. *Top right:* Number of tracks in the tagging jet as selected from the JetSelection algorithm, as used in [9]. *Bottom left:* angle in the transverse plane between the tagging jet direction and reconstructed signal B candidate. *Bottom right:* cone angle ΔR between the tagging jet direction and the signal B candidate direction.

will be compared to the all processes sample.

2.4 Detector description plots

All tracks in the event with at least 10 stereo and 10 axial COT hits or with at last two $R - \Phi$ hits in the SVX have been refitted using the prescription of [12]. The same settings have been used for Monte Carlo and data.

The number of COT hits per track is shown in Fig. 6. Obviously the number of COT hits is not correctly modeled in Monte Carlo. The data distribution is more smeared and the central value is smaller than in the simulation. Both axial and stereo hit numbers are different for data and Monte Carlo.

COT simulation is under development at the moment of writing, so the discrepancy in the number of hits is expected to disappear in the next software releases.

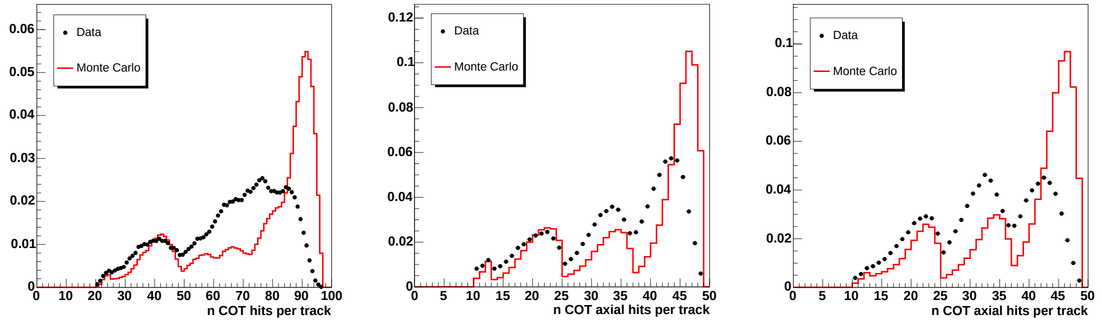

Figure 6: Distributions of the number of COT hits per track for data and Monte Carlo. *Left plot:* all hits. *Center plot:* axial hits. *Right plot:* stereo hits.

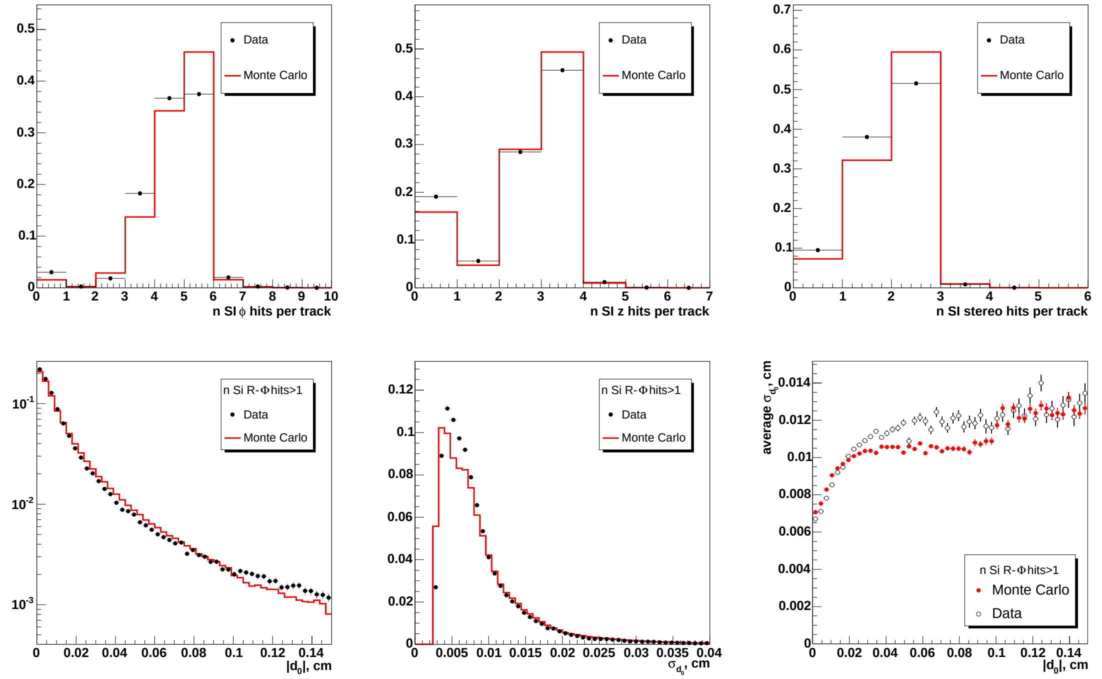

Figure 7: *Top row:* Distribution of the number of hits per track for data and Monte Carlo. *Bottom row:* Track impact parameter corrected with respect to the beam spot (*left*), error on impact parameter corrected by the beamline error (*center*) and their correlation (*right*).

The plots in the top row of Fig. 7 shows the comparison between data and simulation for the number of silicon hits per track.

Tracks in simulated events appear to have in average more z and stereo hits than in the data. The number of Φ hits is in good agreement.

The impact parameter distributions corrected with respect to the beam spot (Fig.7, bottom left plot) present a small discrepancy in the tail. The error on the impact parameter corrected by the beam spot error (Fig.7, bottom center plot) shows a disagreement as well. Moreover one observes that the correlation between impact parameter and its error is not perfectly modeled in the simulation: in the d_0 range of 0.02 cm - 0.1 cm the Monte Carlo resolution is better than the data resolution. A closer look showed that the disagreement is mostly given by tracks that have their innermost Φ hit in layer 1. In general the innermost

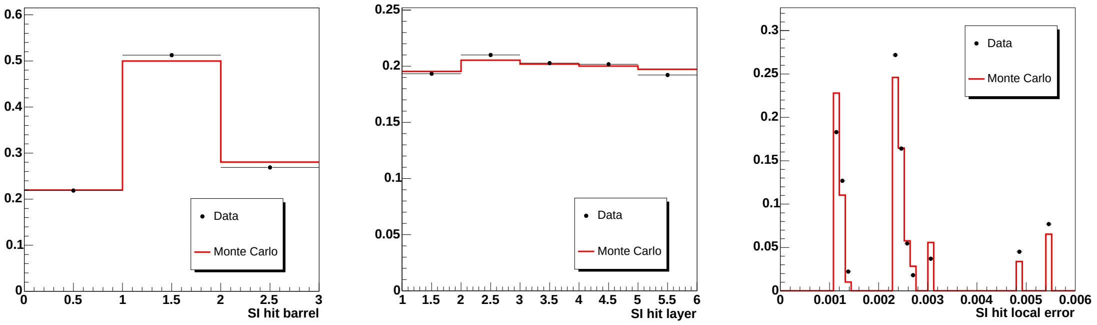

Figure 8: Hit position in the silicon detector barrels and layers, and hit error distributions for data and Monte Carlo.

hit and its error dominate the measurement of the impact parameter of the track and its error.

The effect depends on the position of the track hits in the silicon detector and on the errors associated to the hits. We observe compatibility between data and Monte Carlo concerning the position of the hits in the silicon barrels and in the silicon layers (Fig. 8). The error on the hit position is in average bigger in data than in simulated events. The disagreement is similar for each layer of the silicon detector. The error is wrongly modeled also in the innermost hit, which explains the disagreement seen in d_0 and σ_{d_0} distributions.

3 Event shape study

In this section we present some plots regarding the kinematic of tracks on the same side and on the opposite side of the event. The aim is to understand the clustering algorithms currently in use in the B group.

The definition of same side and opposite side is the same given in section 2.3.

In the following we indicate with "B" any non excited b-hadron. Our definition includes $B^0/\bar{B}^0$ (∼40%), B^+/B^- (∼40%), $B_s/\bar{B}_s$ (∼10%) and $\Lambda_b/\bar{\Lambda}_b$ (∼10%). With B^{**} we intend any excited b-hadron.

We display plots regarding the distributions of the tracks corresponding to particles coming from the B decay, from the B^{**} decays and from the fragmentation of the b quark into hadrons. The identification of such particles, once the b-hadrons have been identified in the OBSP list, is performed in the following way:

- **particles from B decay:** the B is identified in the HEPG list and all the final state particles coming from its decay are stored in a vector. This list of particles includes the daughters of intermediate state particles originating from the B, such as D mesons, J/Ψ, etc.

- **particles from fragmentation:** in Pythia the fragmentation of quarks into hadrons is handled via a *string particle*, which has code 92 in HEPG bank. The fragmentation particle for each B in the event is found by going up in the generation tree in the HEPG bank. The daughters of the string particles are then stored into a vector and considered fragmentation particles. As a caveat we must state that the distinction between fragmentation tracks from signal b and tagging b is only valid when the heavy quarks are fragmented via different strings. The latter seems to be the most frequent case in our Pythia Monte Carlo sample.

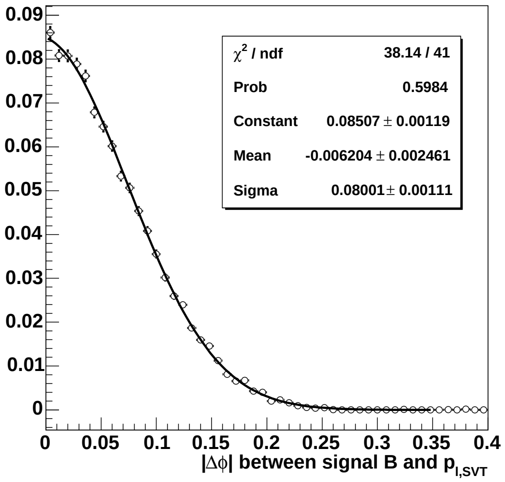

Figure 9: Resolution on the direction of the signal B in the $R - \phi$ plane. A gaussian fit is overlaid to the Monte Carlo points.

- **particles from B^{**} decay:** if one of the string particle daughters is an excited b-hadron, its daughters are stored in a separate vector. This vector does not include the B decay products.

Each track in the event is matched to its HEPG parent. If the parent is in any of the vectors above mentioned, the track is flagged accordingly.

3.1 Same Side

It has already been mentioned that for each event, the signal B is reconstructed as the fit of a vertex with the trigger lepton and the SVT track.
In order to understand how good our resolution is on the signal B direction reconstruction, we look at the angle $|\Delta\Phi|$ in the transverse plane between the reconstructed B direction and the true one (Fig. 9).
A gaussian fit of the angle distribution gives a σ equal to 0.080 ± 0.001 rad. This number can be interpreted as the signal B reconstruction resolution.
Although the signal B is reconstructed out of two tracks only, its direction in the transverse plane is quite precise. The trigger tracks have relatively high momentum (2 GeV and 4 GeV). These tracks are probably the tracks carrying the highest fraction of the B momentum, so the sum of their momenta identifies rather well the B direction.
The properties of the tracks originated from the signal B are displayed in the plots in Fig.10. The tracks coming from the B decay and from the excited B decay follow closely the B direction (fig 10, top left). The plot in Fig 10, top right, is the distribution of the cone angle of the furthest track from the true signal B direction. It shows that the tracks produced in the decay of B** are closer to the B direction than those produced in the B decay.
The fragmentation tracks instead appear to be distributed uniformly in the detector. The behavior does not seem to depend on the transverse momentum of the signal B. The bottom left plot in Fig.10 shows that for increasing B transverse momentum the B decay tracks become more collimated around the mother direction. The average maximum distance of fragmentation tracks from the B direction does not vary significantly.
The ΔR plot indicates that the isolation cut of 0.7, which is also used in [9], rejects the tracks from signal B decay with high efficiency.

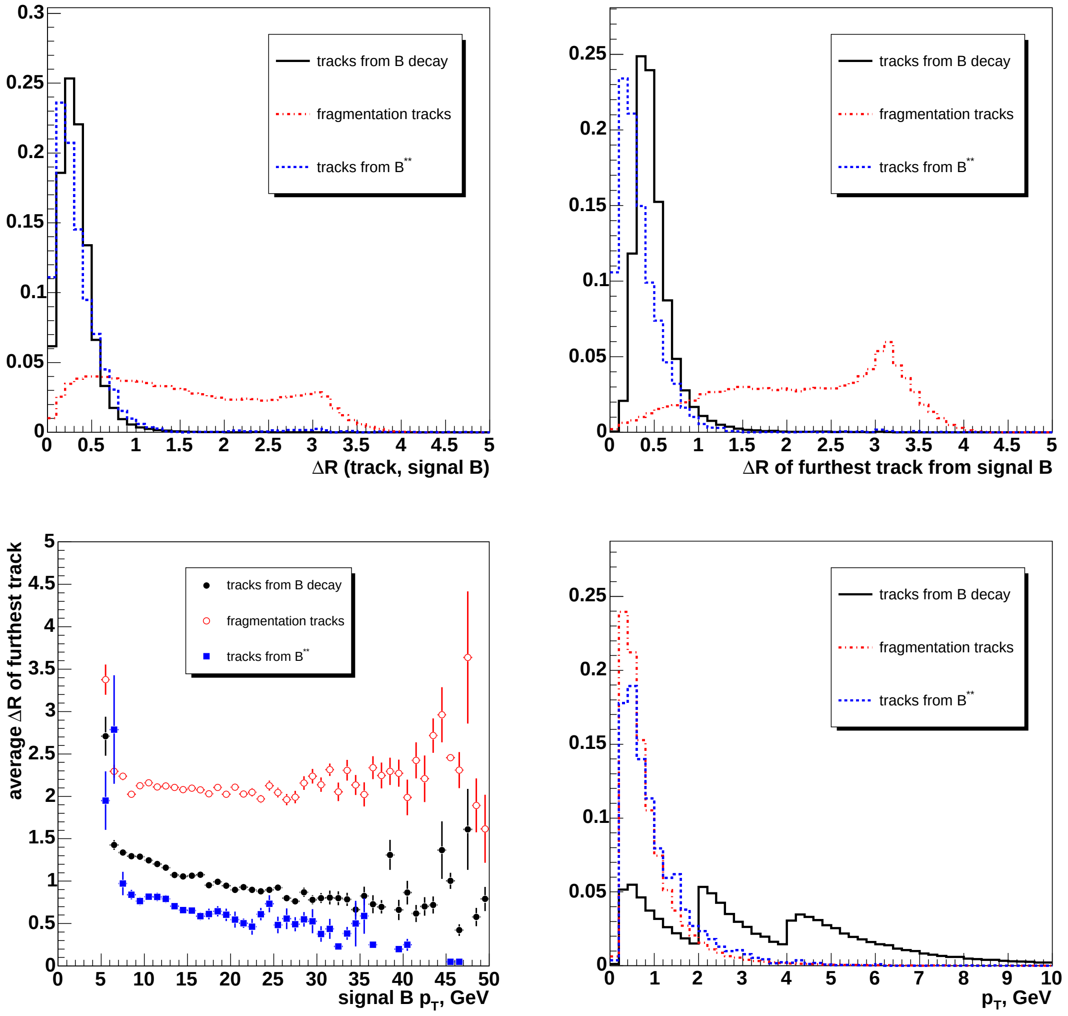

Figure 10: Angular distribution, with respect to the true signal B direction, of tracks originated from the signal B. *Top left plot:* cone angle of all tracks. *Top right:* cone angle of the furthest track from the true signal B direction. *Bottom left:* Dependency on signal B transverse momentum of the cone angle of the furthest track. *Bottom right:* Transverse momentum spectra for tracks coming from B decay, from B^{**} decay and from fragmentation.

The fraction of tracks that fail the isolation cut is shown in Fig. 11. In average, if particles from B or excited B decay are in the detector, 5% of their tracks fail the isolation cut. The average fraction of fragmentation tracks falling outside the isolation cone is 74%.
It does not seem possible to isolate the fragmentation tracks and to define uncorrelated sides of the event. A big portion of fragmentation tracks originated on the signal side is always on the opposite side. Since fragmentation tracks carry flavor information, their presence on the opposite side might spoil the computation of opposite side flavor.
The transverse momentum distribution of tracks on the signal side (Fig. 10, bottom right) shows that the B decay tracks are more energetic than the fragmentation and B$**$ tracks. One notices also that the trigger cuts on momentum applied to the lepton and SVT tracks give a characteristic shape to the distribution of B decay tracks p_T.
The p_T distribution of tracks that fail the isolation cut is shown in Fig. 11 (right plot). The tracks that fail the cut are soft in general, but a fraction of tracks with $p_T > 1$ GeV can also be found on the opposite side. These tracks satisfy then the minimum p_T cut of

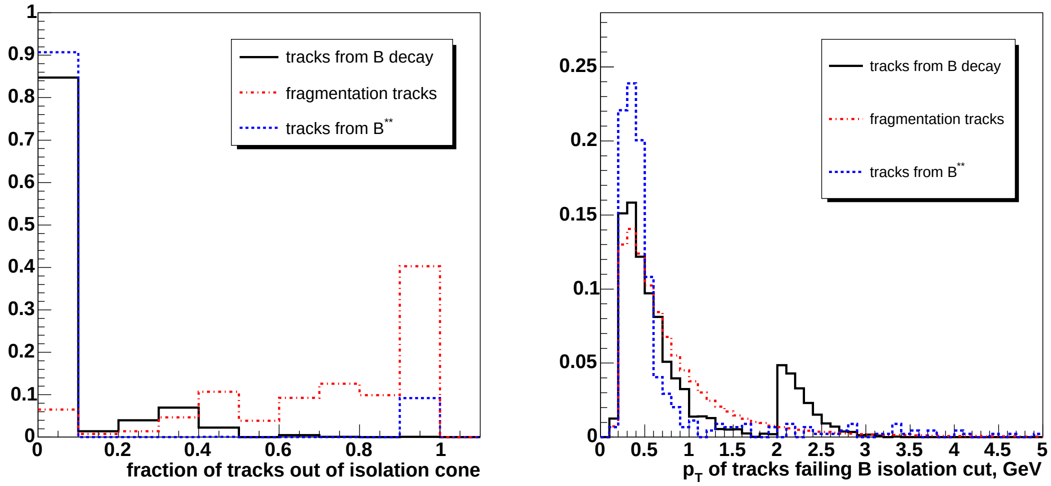

Figure 11: *Left plot:* Fraction of tracks related to the signal B that fail the isolation cut ($\Delta R = 0.7$). *Right plot:* Transverse momentum of the tracks failing the isolation cone.

seeds for Cone Clustering algorithm.

3.2 Opposite Side

The true tagging b quark does not have to satisfy requirement at generator level, in particular it does not have to pass any transverse momentum cut. Consequently the tagging B is in average less energetic than the the signal B. The angular distribution of its decay products reflects this feature (Fig. 12, top left). The tracks coming from the tagging B decay are less collimated than the tracks from the signal B decay. The furthest decay track might also be far away in angle from the true tagging B direction, as the long tail in the ΔR distribution in Fig. 12 (top right) indicates.

A cone of 1.5 around the true tagging B direction collects the most of the B decay tracks and excited B decay tracks, but it includes also a considerable fraction of fragmentation tracks.

The tagging B tracks are softer than the signal B tracks (compare Fig. 12 bottom right with Fig. 10 bottom right). The tracks coming from the tagging B decay have in average higher momentum than the fragmentation tracks and the most of them have p_T smaller than 1 GeV.

As expected, the tracks from B and excited B decay get closer to the true B direction as the heavy hadron transverse momentum grows (Fig. 12, bottom left). The maximum cone angle of the fragmentation tracks does not depend on the p_T of the B.

It is not possible to find a cone around the the B direction which could reject all fragmentation tracks.

In an ideal situation, as for example in the Monte Carlo with flavor creation only, the most probable tagging B direction could be found as the opposite to the direction of the signal B candidate. The tagging B transverse momentum could be estimated from the signal B p_T. A cone clustering algorithm could then use a cone size parametrized with the tagging B p_T. The parametrized cone size would allow the inclusion in the jet of most of the B decay tracks and the reduction of the fragmentation tracks fraction.

The reconstruction of pure jets with a cone algorithm is in reality more complicated.

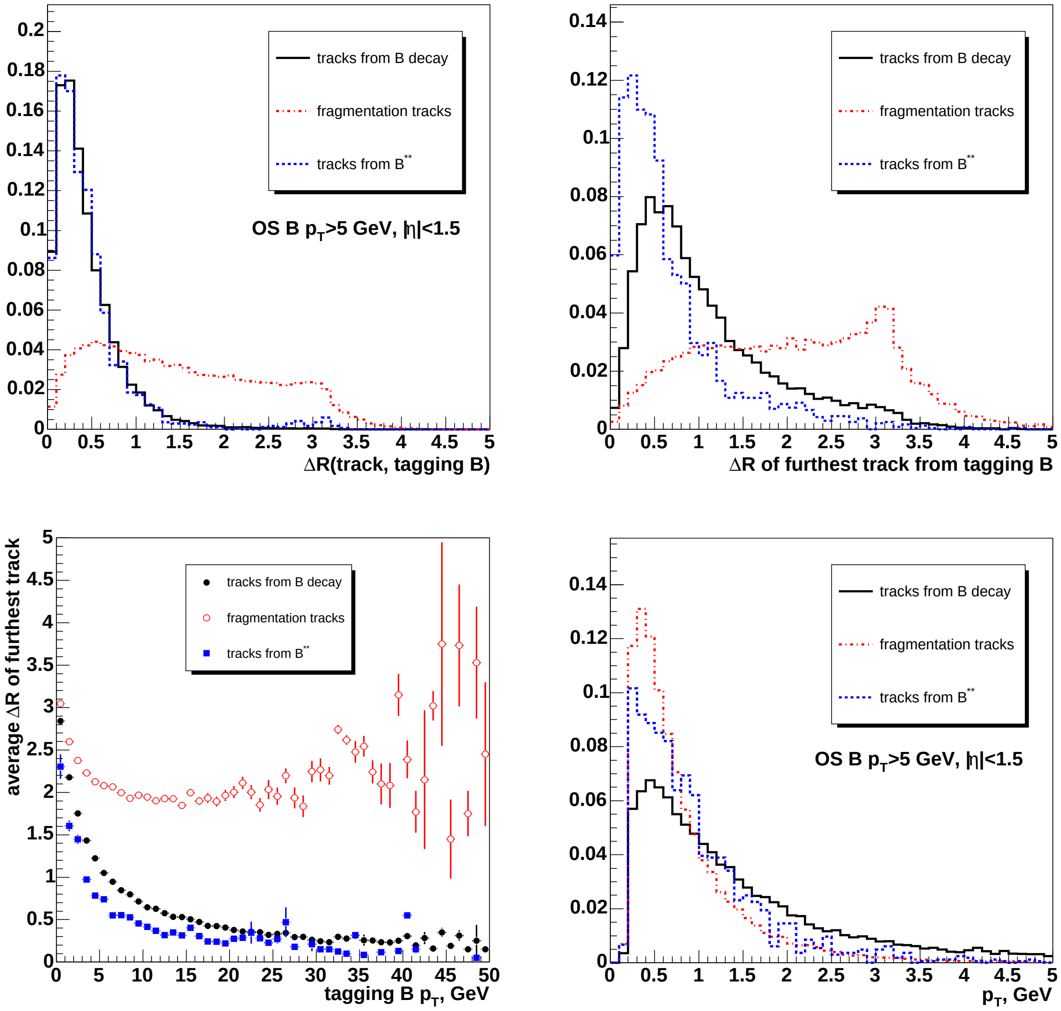

Figure 12: Angular distribution of tracks originated from the tagging B with respect to the true tagging B direction. *Top left plot:* cone angle of all tracks. *Top right:* cone angle of the furthest track from the true tagging B direction. *Bottom left:* Dependency on signal B transverse momentum of the cone angle of the furthest track. *Bottom right:* Transverse momentum spectra for tracks coming from B decay, from B^{**} decay and from fragmentation.

We have learned from the Monte Carlo sample with all processes that the signal and the tagging B directions are weakly correlated. Only in few cases the heavy hadrons fly back to back. Thus we can not use the signal B candidate to find out the tagging B candidate direction.

The correlation between the heavy quarks transverse momenta is also weak, consequently the b-hadrons can have very different p_T. The tagging B p_T should then be estimated via an exclusive or an inclusive reconstruction on the opposite side. This strategy though would be successful only in a small fraction of events.

The scenario suggested by the all processes Monte Carlo is far from being ideal. The reconstruction with the Cone Clustering algorithm of jets containing most of the B decay products and the least of fragmentation tracks proves to be a challenging task.

A different approach in jet reconstruction is offered by Mass Clustering, described in [2].

The invariant mass of all tracks on the opposite side is distributed as in Fig 13 (left plot). The most probable value for the mass is rather high, around 15 GeV. The invariant mass

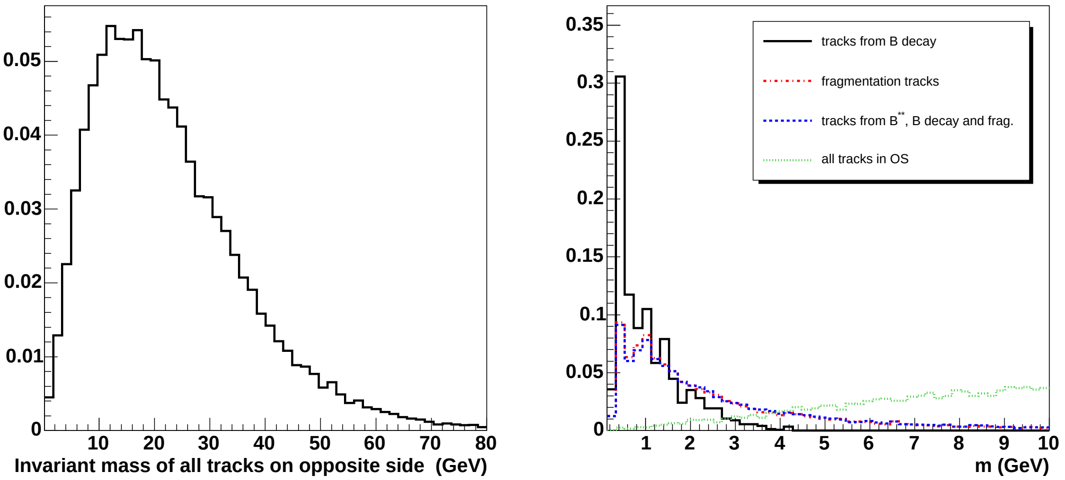

Figure 13: *Left plot:* Invariant mass distribution of all tracks on the opposite side. *Right plot:* invariant mass of tracks coming from B decay and/or from fragmentation, and comparison to the invariant mass of all tracks on the opposite side.

of the tracks coming from the B decay is shown in Fig.13(right plot). The mass is always smaller than 5 GeV and it peaks evidently at low values when there is only one particle from B decay in the detector. The invariant mass of fragmentation tracks is in average higher than the mass of the B decay tracks and it is almost always below 10 GeV.
The mass of all tracks on the opposite side is significantly higher than the mass of fragmentation tracks and B decay products.
The mass distributions indicate that a jet containing a very high fraction of B decay products should not have an invariant mass higher than 5 GeV.
The reconstruction of the jet with a cutoff on the minimum invariant mass is rather close to a cone based reconstruction. High p_T jets have a smaller cone size and lower p_T jets are broader. Fig. 12 (lower left plot) suggests that this strategy is the ideal one. The algorithm could be tuned to give jets containing mostly B decay tracks.

4 Clustering algorithms evaluation

4.1 Cone Algorithm

In order to understand which is an optimal value for the cone size used by the Cone Clustering algorithm, we performed a scan of the cone angle parameter. We evaluated how good the algorithm is in collecting tracks from B decay and from fragmentation.
The cone angle values used for this test were 0.4, 0.7, 1 and 1.5. The other parameters were set to the optimized values given for the blessed Jet Charge Algorithm [9].
We define the *purity* of a jet is defined as the number of tracks from B decay (or fragmentation tracks) in the jet divided by number of tracks in the jet.
The *efficiency* is defined as the number of tracks from B decay (fragmentation tracks) in the jet divided by the number of B decay tracks (fragmentation tracks) on the opposite side.
The *best jet* in the event is the reconstructed jet that contains the highest number of tracks from the tagging B decay. If there are two jets with a similar number of B decay tracks,

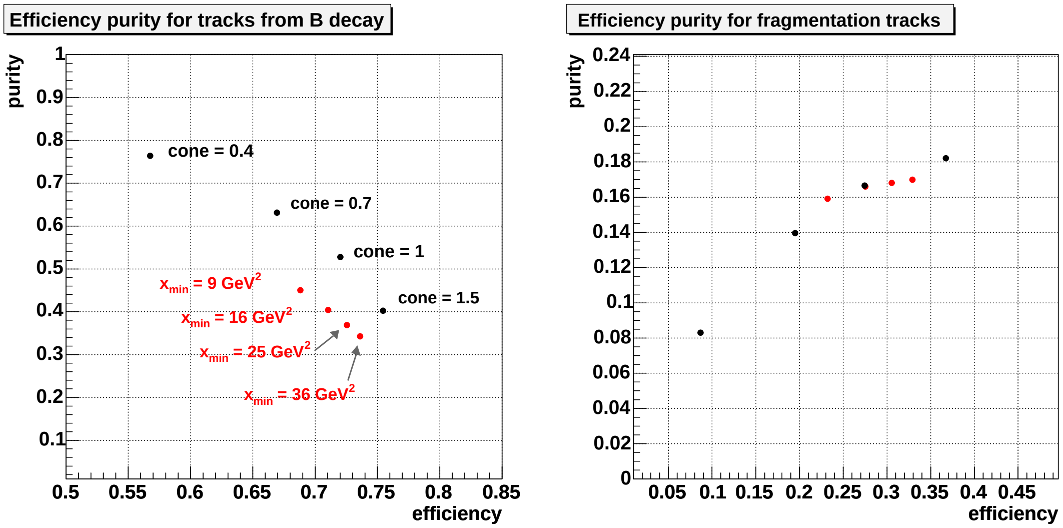

Figure 14: Evaluation of efficiency and purity for cone (black points) and mass algorithms (red points) for different cone values and x_{min} values.

the best jet is the closest one to the true tagging b-hadron direction. In events in which all B decay products are outside the tracking detector, there is no best jet.

The efficiency and purity of the algorithm for a given cone size are then the average efficiency and the average purity of the best jets that the algorithm reconstructs.

Plots of the performance of the Cone Clustering algorithm for different cone angle values can be seen in Fig. 14 (black dots). As expected, the highest efficiency of collecting B decay tracks is given by the largest cone, the highest purity is given by the smallest cone (Fig. 14, left plot). The reconstructed jets with a cone of 0.4 consist mostly of a single track, so although they are rather pure jets they are not the best choice for flavor tagging purposes. As the cone size grows larger, more fragmentation tracks are included in the jet (Fig. 14, right plot). The growth of fragmentation track purity indicates that, by increasing the cone size, more fragmentation tracks constitute the jet, but not more tracks from B decay.

A hint on the tagging B direction resolution of the algorithm is given by the angle in the transverse plane between the best reconstructed jet and the true tagging B direction. A distribution of this quantity for each cone size is shown in Fig. 15 (left plot). Only events in which the tagging B is in the acceptance and has a transverse momentum larger than 5 GeV are used to produce the resolution plot.

The most peaked distributions are obtained with the cone sizes 0.7 and 1. Thus we conclude that the cone sizes 0.7 and 1 give a better resolution on the tagging B direction than very small cones ($\Delta R = 0.4$) and very large cones ($\Delta R = 1.5$). This conclusion agrees with the plot in Fig. 12 (top left). In a cone of size between 0.7 and 1 around the true B direction there are ~90% of the B decay tracks and ~30% of the fragmentation tracks.

4.2 Mass Algorithm

We performed a simple scan of the invariant mass cutoff x_{min} to evaluate the efficiency and purity of the Mass Clustering algorithm. The chosen values were 9 GeV^2, 16 GeV^2, 25 GeV^2 and 36 GeV^2. The minimum jet mass was the only parameter varied, all the other parameters were taken with their default values.

The definition of purity, efficiency and best jet for the mass clustering algorithm are the

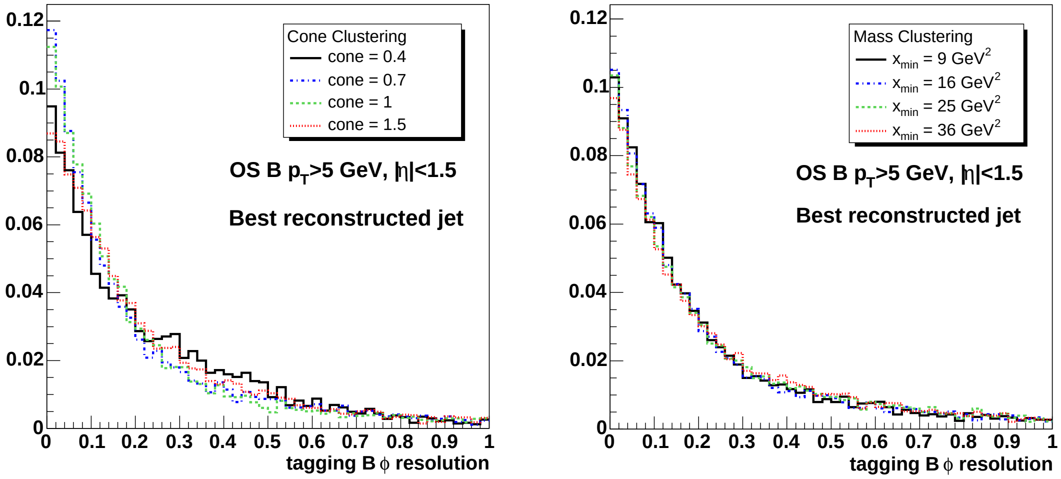

Figure 15: Angle in the transverse plane between the true tagging B direction and the best jet reconstructed by the cone clustering algorithm (*left plot*) and by the mass clustering algorithm (*right plot*) in events in which the tagging B is in the detector acceptance.

same given for the cone algorithm (see section 4.1).
The performance of the algorithm for different x_{min} values is shown in Fig. 14 (red dots). In average the mass algorithm is less pure than the cone algorithm. Less than half of the tracks in the best jet are B decay tracks.
The fraction of fragmentation tracks included in the jet increases as the jet size becomes larger (Fig. 14, right plot), although the variation in purity and efficiency is over a small range.
The plot on the right in Fig.15 shows the angle in the transverse plane between the true tagging B direction and the best jet reconstructed by Mass Clustering algorithm. A cut is applied to select events in which the tagging B is in the acceptance and has a transverse momentum larger than 5 GeV. The distributions corresponding to different x_{min} cut do not show significant differences. This indicates that x_{min} is not a parameter of primary importance in the tuning of Mass Clustering algorithm. Other parameters, like the minimum transverse momentum of the seed tracks, might have a more significant effect on the resolution of the algorithm.
It should be noticed that the resolution of the mass algorithm is worse than the resolution obtained by Cone Clustering algorithm using the cone sizes 1 and 0.7, and is slightly better than the resolution of cone 0.4 and 1.5.
Although the Mass Clustering algorithm has a stronger physical motivation than the Cone Clustering algorithm, its performance with the default set of parameters is not better.
A more accurate and wide scan of parameters should indicate the optimal configuration to run Mass Clustering.
At the moment Cone Clustering algorithm with a cone of 0.7 or 1 should be preferred to Mass Clustering to reconstruct jets on the opposite side of ℓ+SVT events.

5 Jet Selection algorithm evaluation

In this section we use the all processes Monte Carlo sample to evaluate the purity of algorithm to select jets as it is used in [2].
The JetSelection algorithm chooses the tagging jet in an event out of the reconstructed jets

	Selected jets	Best jets	Purity
SecVtx algorithm	5780	4400	76%
min I.P. jet probability	9600	5260	55%
Highest p_T	32500	10600	33%

Table 1: Purity of JetSelection algorithm setup like in [2].

on the opposite side by the Cone Clustering algorithm.

If a jet on the opposite side has a secondary vertex found by SecVtx algorithm [13] with transverse decay length significance greater than 3 then it is selected. If none of the jets has a secondary vertex the jet with the smallest impact parameter based probability is chosen, provided that the probability is smaller than 0.1. If the jet probability algorithm is not successful, the jet with the highest transverse momentum is selected.

The selection algorithm is structured in a way that always allow to select a tagging jet, as long as at least one jet is reconstructed by Cone Clustering.

We define the best jet in the event as in section 4.1.

The *purity* of the algorithm is the number of best jets selected divided by the number of all jets selected. Table 1 shows for each JetSelection method the number of jets selected, the number of best jets selected and the purity of the jet sample.

As expected SecVtx algorithm provides the highest purity sample and it is able to tag a small number of jets (about 12% of all jet chosen by JetSelection).

About 20% of all selected jets is tagged by the minimum probability method. These jets have a purity of 55%.

The majority of the tagged jets is chosen by highest p_T and the purity of this sample is rather low, only 33%.

The purity hierarchy of the three methods explains the different tagging power of the jet samples measured by Jet Charge Tagger [2].

We have seen in section 2.2 that the tagging B is outside the detector in a high fraction of events. Since in these events the B decay products can not be reconstructed, none of the opposite side jets should be selected as a tagging jet. An improvement to JetSelection could be a method alternative to the highest p_T. This method should avoid to choose any jet in events without B decay products in the tracking detector. The selected jets would then constitute a purer sample and would contribute with higher tagging power to Jet Charge Tagger.

6 Conclusion

We have proved in this note that a Monte Carlo sample containing only $b\bar{b}$ flavor creation processes does not correctly describe the data. Such a Monte Carlo sample might lead to too optimistic efficiency of identifying the tagging b-hadron.

We strongly recommend to develop opposite side tagging algorithms on a Monte Carlo containing all $b\bar{b}$ production processes, as we have shown that such a Monte Carlo sample reproduces the distributions of opposite quantities as seen in data. Although a fine tuning of Pythia fragmentation parameter is needed, the all processes Monte Carlo produced with default parameters describes the data very well.

We have used the all processes Monte Carlo to estimate the purity of the opposite side clustering algorithms based on Cone and Mass Clustering. A simple scan of the cone angle and the mass cutoff, respectively, led to the conclusion that the Cone Clustering algorithm with ΔR 0.7 or 1 gives the best result in term of resolution on the tagging B direction for the lepton+SVT sample.

The Monte Carlo sample has also been used to compute the purity of jet samples selected by three different JetSelection methods. The computed purities justify the tagging power that each jet sample brings to Jet Charge Tagger. JetSelection always finds a tagging jet if at least a jet is reconstructed on the opposite side. According to our generator level study, this feature reduces the purity of the sample of selected jets.

References

[1] T. Sjöstrand, Comput. Phys. Commun. **82** (1994) 74

[2] G. Bauer et al., *"Improved Jet Charge Tagger for summer conferences 2004"*, CDF note 7131

[3] E. Norrbin and T. Sjöstrand, *"Production and hadronization of heavy quarks"*, hep-ph/0005110

[4] R. Field, *"The Sources of b-Quarks at the Tevatron and their Correlations"*, CDF note 5813

[5] J. D. Lewis, P. Avery, *"CLEOMC: The CDF interface to the CLEO Monte Carlo (QQ)"*, CDF note 2724

[6] http://www-cdf.fnal.gov/internal/physics/bottom/b-montecarlo/db/g020.txt

[7] W. Bell, J.P. Fernandez, L. Flores, F. Wuerthwein, R.J. Tesarek, *"User Guide For EvtGen @ CDF"*, CDF note 5618

[8] M. Jones et al., *"Sample Composition of the ℓ+SVT Triggers"*, CDF note 6480

[9] G. Bauer et al., *"B Flavor Tagging Using Opposite Side Jet Charge"*, CDF note 6951

[10] http://cdfkits.fnal.gov/CdfCode/source/JetUserObjects/JetUserObjects/ConeClusteringAlg.hh

[11] http://cdfkits.fnal.gov/CdfCode/source/BottomTaggers/BottomTaggers/JetSelectionAlg.hh

[12] M. Campanelli, E. Gerchtein, *"Calibration of the momentum scale for Kalman refitter using J/ψ events"*, CDF note 6905

[13] http://cdfkits.fnal.gov/CdfCode/source/BTagAlgs/BTagAlgs/SecVtxAlg.hh