Summerstudent Program 2009, DESY

Top quark analysis project at CMS

Juri Smirnov[a]

[a] University of Munich

September 9, 2009

Abstract

In this project the detector effect on the measurement of the top quark kinematic variables is studied. The binning of the histograms for the different kinematic variables is optimized with the goal to minimize the migration effects from one bin into another. A technique for a quantitative analysis of the binning's validity was developed and with its help the bin sizes were optimized.

It was discovered that the kinematic variables for the top quark are well reconstructed and an optimal binning can be found. The situation is different for the kinematic variables of the top antitop system. Here systematic migration is observed and does not allow to use the bin to bin correction technique for the deconvolution of the detector. Nevertheless there was an effort made to reduce the systematic migration for the angular variables of the top antitop system. It will be shown that in some cases this reduction can be made.

Contents

1 Introduction **2**

1.1 The CMS detector 2

1.2 The top quark 2

1.2.1 The top quark production 3

1.2.2 The top decay 4

2 The kinematic variables **6**

2.1 The top kinematic variables 6

2.2 The t t-bar kinematic variables 7

3 The effect of the detector and the reconstruction **9**

3.1 Estimation of the detector's visible range 10

3.1.1 The top kinematic variables 10

3.1.2 The t t-bar kinematic variables 11

3.2 The quantification of the migration and bin size adaption 11

3.2.1 The top kinematic variables 14

3.2.2 The t t-bar kinematic variables 14

3.2.3 The t t-bar Phi problem 15

4 Further studies **16**

4.1 The comparison of the hadronic/leptonic decayed top 16

4.2 The improvement of the t t-bar Eta resolution 17

4.3 The application of a realistic event hypothesis 18

4.3.1 The stability and purity of the top quark with the kGeom algorithm 18

4.3.2 The improvement of the correlation of the t t-bar Eta applying the kGeom algorithm 19

4.4 Suggestions for the t t-bar Pt reconstruction 19

5 Conclusion **21**

6 Acknowledgments **22**

1 Introduction

1.1 The CMS detector

The CMS detector continues a long tradition of barrel detectors. Since the design is very classy there is a short listing of the main components only. Going from the center to the outer parts the detector contains a full silicon tracker, a scintillating electromagnetic calorimeter of PbW04, a hadron calorimeter , a superconducting solenoid which produces a 4 Tesla magnetic field inside it and a 2 T field in the iron yoke. The myon chambers are very large and powerfull in detecting high energetic muons. The interesting feature is that the h. cal is completely in the solenoid. This makes the detector very compact , but might also cause problems since some of the jet energy can be lost when the hadrons energy is very high and it penetrates the rather short calorimeter. An other ineresting feature is the huge magnetic field which enables the detector to measure the momentum of high energetic muons.

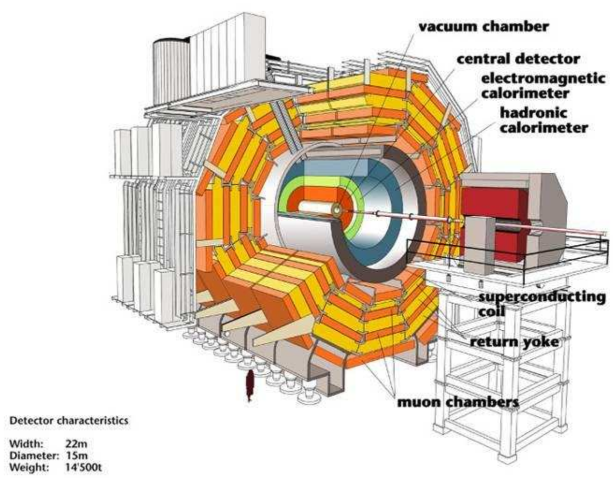


Figure 1: CMS detector

1.2 The top quark

With the mass of (171.2±2.1)$\frac{GeV}{c^2}$ the top quark is the heaviest quark of the standard model. It has been discovered in 1995 at the Fermilab. The top quark interacts primarily by the strong interaction but can only decay through the weak force. It almost exclusively decays to a W boson and a bottom quark. The Standard Model predicts its lifetime to be roughly $5 \cdot 10^{-25}s$. This is about 20 times shorter than the timescale for strong interactions, and therefore it does not hadronize, giving physicists a unique opportunity to study a "bare" quark of which the spin and the polarization of can be measured. The charge of the top is not measured very precisely so far since the collected number of events at the Tevatron is not high. For instance the hypothesis of the top quark's charge being $\frac{4}{3}$ is excluded by 94% only. Other important quantities for the standard model are the top's lifetime and the CKM transition element. Which is important to prove the CKM unitarity. The precise knowledge of the top quark's mass is an important parameter for the higgs mass prediction. Hence the higgs couples to mass the top loop influences its mass significantly. There is also big interest in the top quark beyond the standard model. Because of its large mass and the coupling to the higgs the top is assumed to be significantly involved in the higgs production process almost independent of the model. On the other hand there are Kalusa Klein theories including extra dimensions. Standard model gauge bosons at high energies can be excited into these dimensions. When those particles, usually called Z', decay they also couple very likely to mass and top events can be observed as the result of such decay channels.

1.2.1 The top quark production

Top quarks are mostly produced in top antitop pairs. Single top quark production was observed in 2006 at the Fermilab as well, but this process is very rare and lies beyond the scope of this project.

At a proton proton collider there are two dominant processes for the top antitop production. The first is light quark antiquark going into a top antitop pair. The second two colliding gluons produce a top antitop pair.

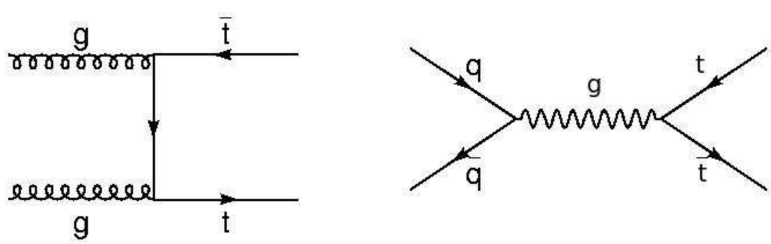


Figure 2: Top production mechanisms

For the t t-bar production Q^2 is above $342^2 GeV^2 \cong 170.000\, GeV^2$. For instance the energy the LHC aims to achieve till the end of the year is 7 TeV, this means 3.5 TeV per beam. It would be sufficient that a parton carries $\frac{1}{20}$ of the proton's momentum to allow a t t-bar production. When one calculates the number of particles which carry at least $\frac{1}{20}$ of the momentum the number of gluons exceeds the quark number by orders of magnitude. Integrating the partial density functions from $x = 0$ to $x = \frac{1}{20}$ it is obvious that the predominant process at the LHC will be the gluon gluon fusion. Since the gluon density functions have not been studied in great detail and their x dependence is not calculatable via perturbation theory this field is of grate interest for physics.

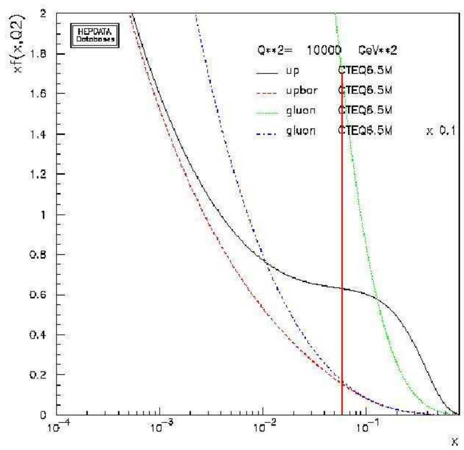


Figure 3: The parton density functions

1.2.2 The top decay

The top quark decays almost exclusively into bottom quarks and radiates a W boson. Since our system contains a top and an antitop it will decay into a bottom and antibottom quark and two $W^{+/-}$bosons. The bottom quarks will hadronize and appear in the detector as jets. The W-boson can decay leptonically or hadronically into a lepton and a neutrino or in two quarks. Here is a table of the possible decay channels for a top antitop pair.

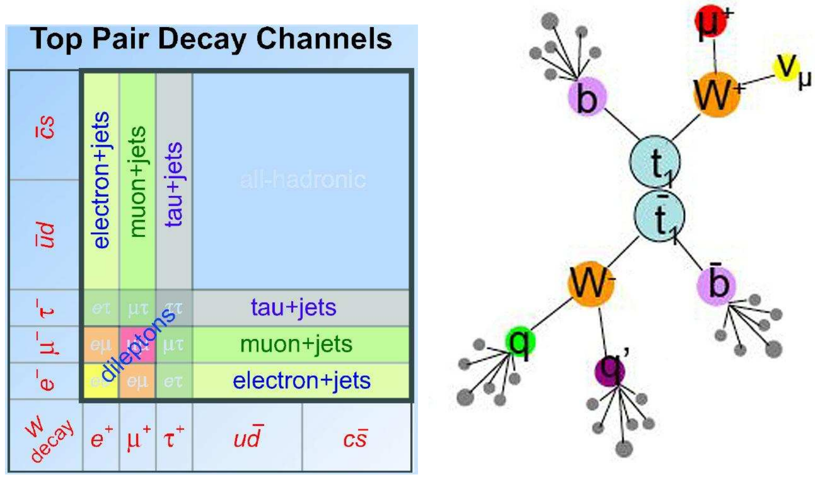


Figure 4: The top decay

- The full hadronic channel has the largest stastics since its phase space with all the color combinations is large, but it also has a large jet background and the reconstruction is difficult because there is a lot of space for combinatorial miss calculations.
- The dileptonic channel has very little background but the statistic is very low and the two outgoing neutrinos make the reconstruction of the event impossible.
- The so called "Golden channel" is the semi-leptonic one. There are sufficient events happening and the background from QCD and W events can be reduced efficiently. Since here only one neutrino goes out it can be reconstructed as missing transverse energy, assuming the detector is hermetic, and the event can be fully reconstructed.

In this project the muonic brunch of the semi-leptonic decay will be of particular interest. This channel is the most promising to study if one aims to calculate the top quark differential crossection. For this purpose the top quarks kinematic variables will be studied and also those ot the t t-bar system. Here is a list of cuts to reduce the different backgrounds.

	$t\bar{t}$+jets s.l. μ	$t\bar{t}$+jets other	Single top s-Ch.	t-Ch.	tW	W+jets	Z+jets	VV+jets	QCD
AllEvents	1,220	7,060	32	832	580	912,000	76,240	236	2,546,279
Trigger	978	1,418	10	260	147	168,633	20,952	100	2,032,021
$\geq$ 1 tight μ	620	345	5	140	69	110,509	15,296	73	7,200
$<$ 2 tight μ	620	309	5	140	66	110,509	9,300	62	7,200
no tight e	620	264	5	140	62	110,508	9,292	53	7,200
veto on loose μ	618	228	5	140	60	110,503	5,492	44	7,192
veto no loose e	616	183	5	140	56	110,469	5,415	34	7,188
$\geq$ 1 jet	614	180	4	125	55	16,998	1,325	18	2,701
$\geq$ 2 jets	593	158	3	63	47	3,076	256	5	387
$\geq$ 3 jets	489	99	1	18	27	651	51	1	60
$\geq$ 4 jets	277	43	0	5	9	140	10	0	7

Figure 5: The cuts

Here a good distinguishing property is the so called tight muon. It means that there is a muon not surrounded by jets. The veto on the loose muon suppresses the di-leptonic channel since one of the separated muons might not fulfill the strong criteria of the tight muon and so fake the semi-leptonic event. The requirement of at least four jets suppresses the QCD background and an other requirement of at least one bottom jet throws out the W+jets background. The bottom quark is the only quark which flavor can be reconstructed from the jet. But the description of the b tagging algorithm is beyond the scope of this work.

2 The kinematic variables

2.1 The top kinematic variables

At first the kinematic variables of a top or antitop quark were studied. The variables of interest here were:

- the radial angle ϕ
- the top's transverse momentum Pt
- and the top's pseudorapidity η

The pseudorapidity Eta needs more explanation at this point. In this context it is not a terrorist organization but a kinematic variable which measures the angle θ of the outgoing particle and the beam pipe in the vertical plane. And the most important feature of it is that it is a good approximation in the high energy limit for the rapidity which is a lorenz invariant value. So for a particle for which the energy exceeds its rest mass η is lorenz invariant. The exact relation between the pseudorapidity and θ is as follows: $\eta = \arctan(\frac{\theta}{2})$

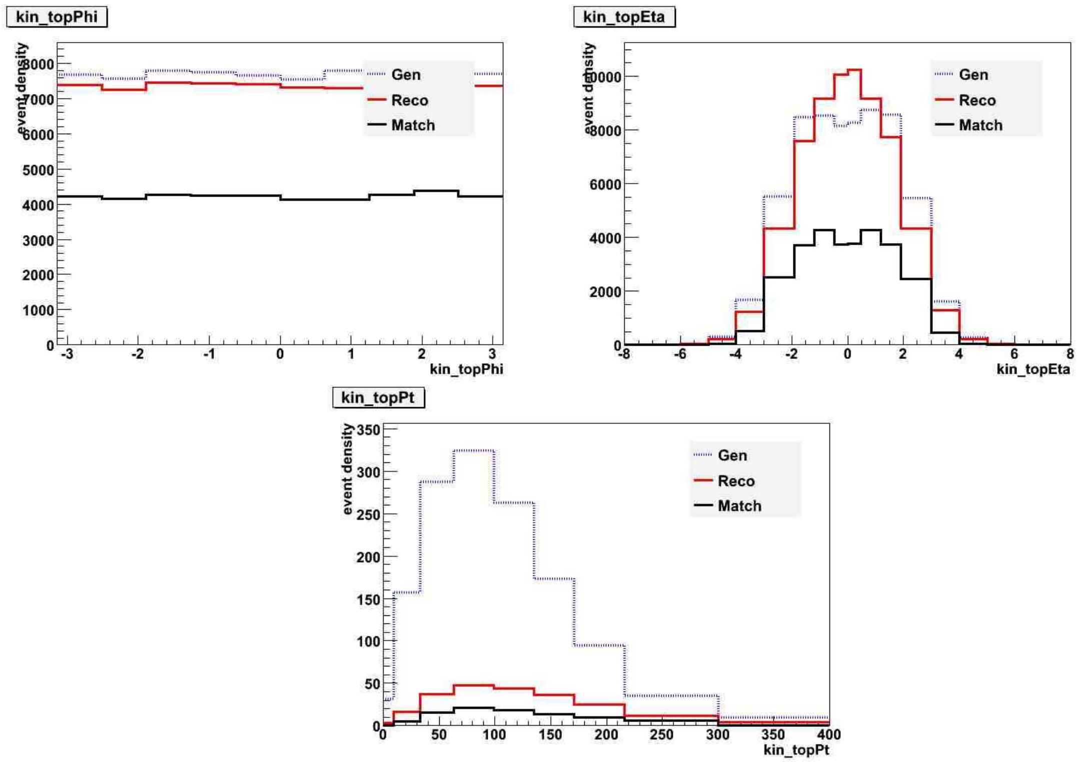


Figure 6: The top kinematic variables

In each plot are three histograms denoted by Gen, Rec and Match. Since in the moment there are no data from the LHC in this project only MC data were studied. So the data generated via MC simulations from the standard model are called Gen. Than those are reconstructed and event hypothesis are produced. Then a testing algorithm checks whether the values are reconstructed with the right property, here it is the Phi angle, within the bin accuracy. So the matching depends strongly on the bin size. The matching will be discussed in detail later on.

The Physics is invariant under ϕ rotations. So the Phi distribution of the top looks flat as expected.

2.2 The t t-bar kinematic variables

An other system of interest is the system of the top and the antitop and its lorenz vector. It can be reconstructed from the Jets and the muon as well but of cause this procedure requires more accuracy. The kinematic variables which can be studied here are:

- Pt of the t t-bar system
- t t-bar Mass
- t t-bar η
- t t-bar ϕ
- t t-bar Delta Phi, which is ϕ(t)-ϕ(t-bar)

Again just the distributions on the generator and reconstruction level are plotted below.

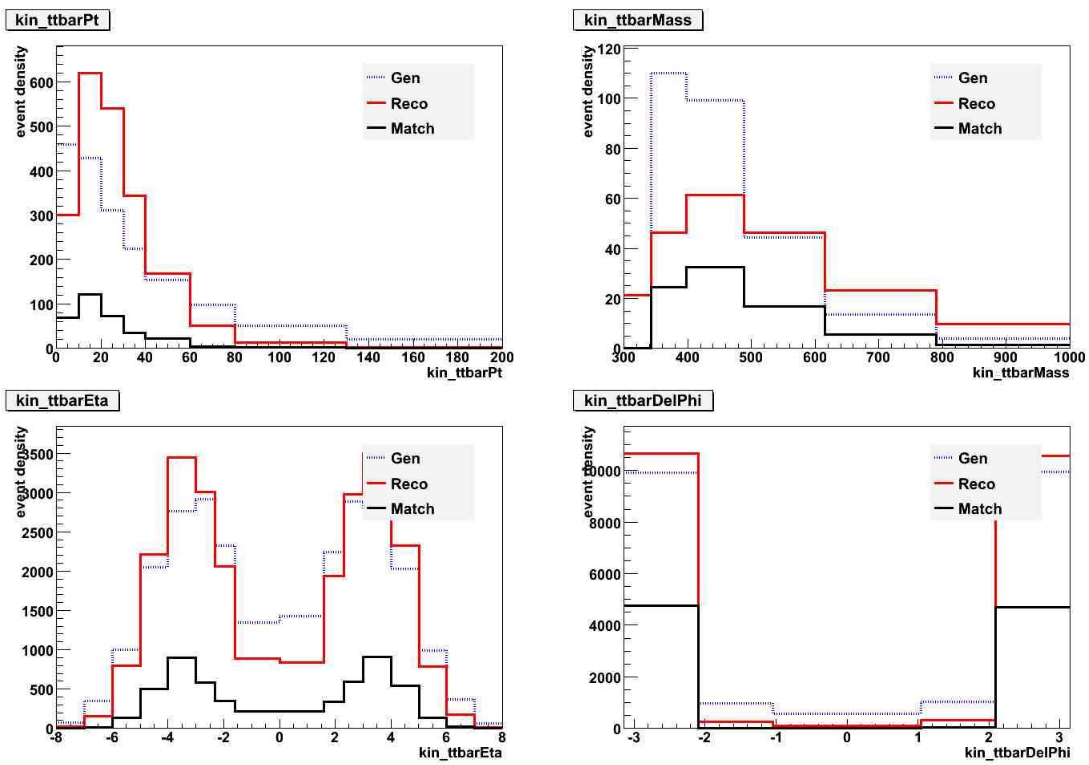


Figure 7: The t t-bar kinematic variables

Again the Phi distribution itself is not of grate interest but the Delta Phi distribution shows, that the t t-bar system is produced predominantely back to back, which on the other hand could be expected. The more interesting feature which can be observed here is that the Eta distribution is not peaked around one. Its maxima are at |Eta|=3 values. As a rule of thumb one can estimate that one unit deviation from zero in Eta corresponds to one order of magnitude in momentum difference between the colliding gluons. This is a feature which is worth studying and should gain a lot of attention to avoid systematic errors which might influence the shape of the distribution.

3 The effect of the detector and the reconstruction

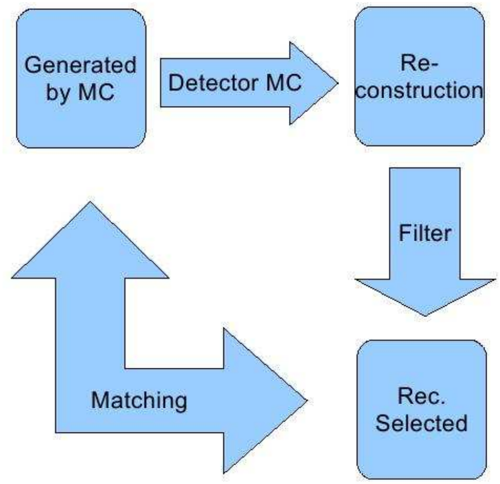


Figure 8: From the generation to the reconstruction

This cycle demonstrates the procedure of data reconstruction in the case of Monte Carlo simulated data. The events are produced via MC simulation according to the Standard Model which is the state of the art in the moment. Than the interaction of the created particles with the detector itself is simulated and one gets data which look like real event signatures in the detector. Than a filter requiring 4 jets one tight muon and so on is applied to select the semileptonic events and finally the reconstructed and selected events are compared binwise to the generated events. So if for instance the Eta of a reconstructed top is within the same bin as of the generated one than this event is matched properly and the matching histogram gets an entry in the corresponding bin. At this point three values will be introduced for analyzing the detector effect on the data and extrapolate later from the measurements to the reality.

- Efficiency: $Eff. = \frac{Rec.Selected}{Generated}$ the reciprocal of the efficiency in the bin gives the correction factor from the number of measured events with a certain property within this bin to the number of real events with this property which did actually take place in the detector. This method though is valid only in the case that there is no systematic migration of the events from one bin into an other

- Stability: $Stab. = \frac{Matched}{Reconstructed}$ this value indicates the systematic migration into a bin. I.E. when there are way more events reconstructed in a bin than matched the excrescent events must migrated into this bin From its neighbours. So a low Stability indicates many events enetering the bin from outside

- Purity: $Pur. = \frac{Matched}{Generated}$ the Purity measurs the opposite case. When there were more events generated in the bin than finally reconstructed, then events left into the neighboring bins. So a low Purity indicates migration out of the bin.

At this point it is important to mention that the migration might be Gaussian distributed. This is the friendly case, when it is possible to use bin to bin corrections if the size is sufficient. In the case when the migration is biasing this becomes rather impossible. Or requires preprocessing of the data.

3.1 Estimation of the detector's visible range

The visible range of the detector was estimated by setting the requirement on the efficiency to be above 10%.

3.1.1 The top kinematic variables

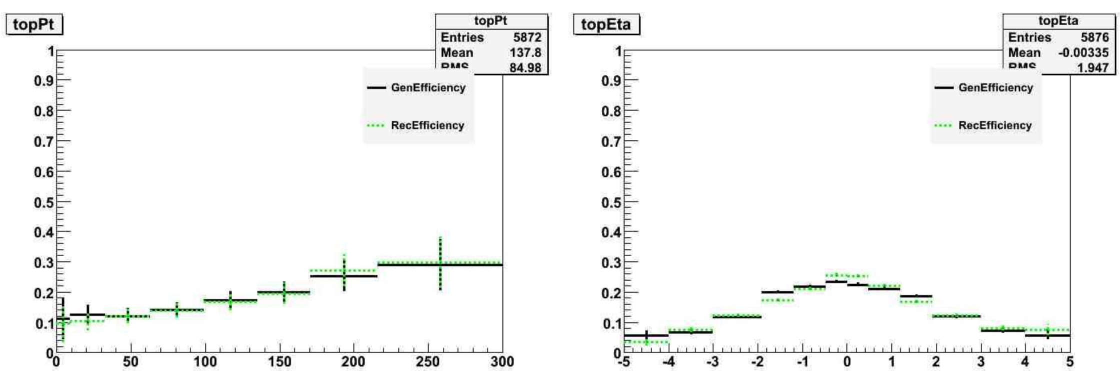


Figure 9: The top efficiencies

The visible range for the top quark can be determined from the plots above to:

- top Pt: Pt > 10 GeV
- top Eta : -3 >Eta > 3

This choice lets the efficiency remain above 10% for the bins in the visible range.

3.1.2 The t t-bar kinematic variables

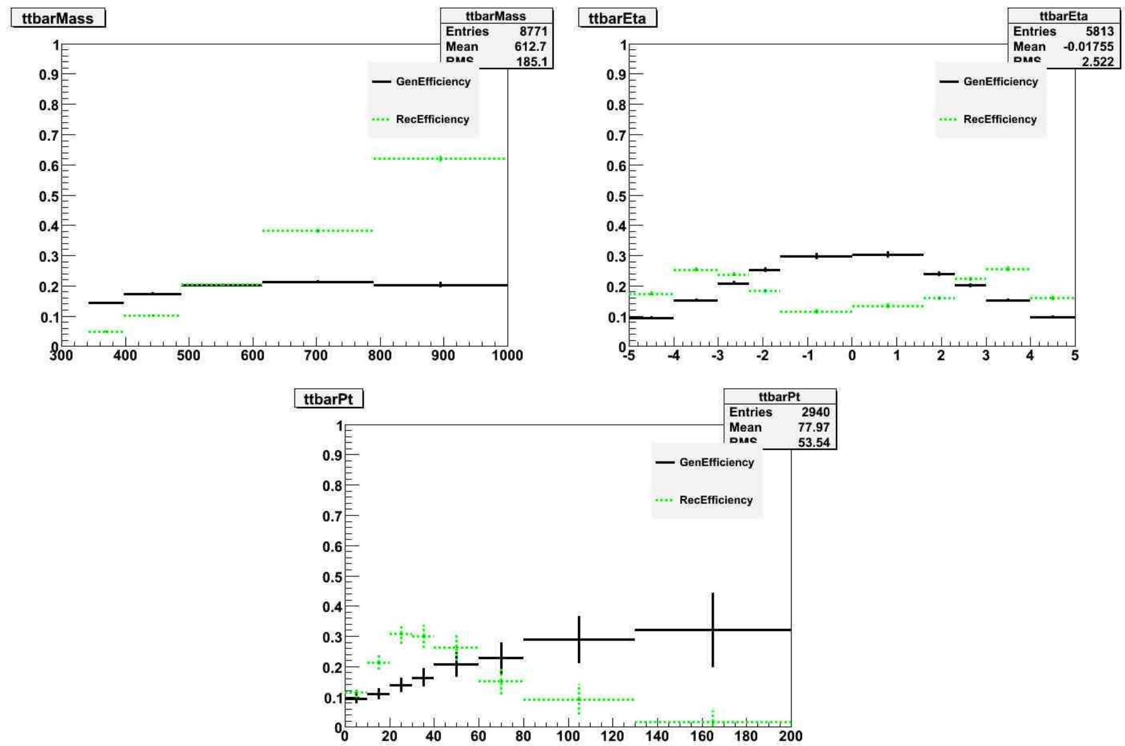


Figure 10: The t t-bar kinematic variables

The visible range for the t t-bar system is more difficult to find, since from the efficiency plots it is already obvious that there will be effects of systematic migration. But this will be discussed later on.

The suggested visible Ranges:

- t t-bar Pt: Pt >10 GeV
- t t-bar Mass: Mass > 340 GeV
- t t-bar Eta: -3> Eta >3

3.2 The quantification of the migration and bin size adaption

In the two dimensional chart the Generated and the Reconstructed data are plotted. Basically this is the representation of the covariance matrix of the two distributions Gen and Rec. In the figure below the Pt value of the top quark is plotted.

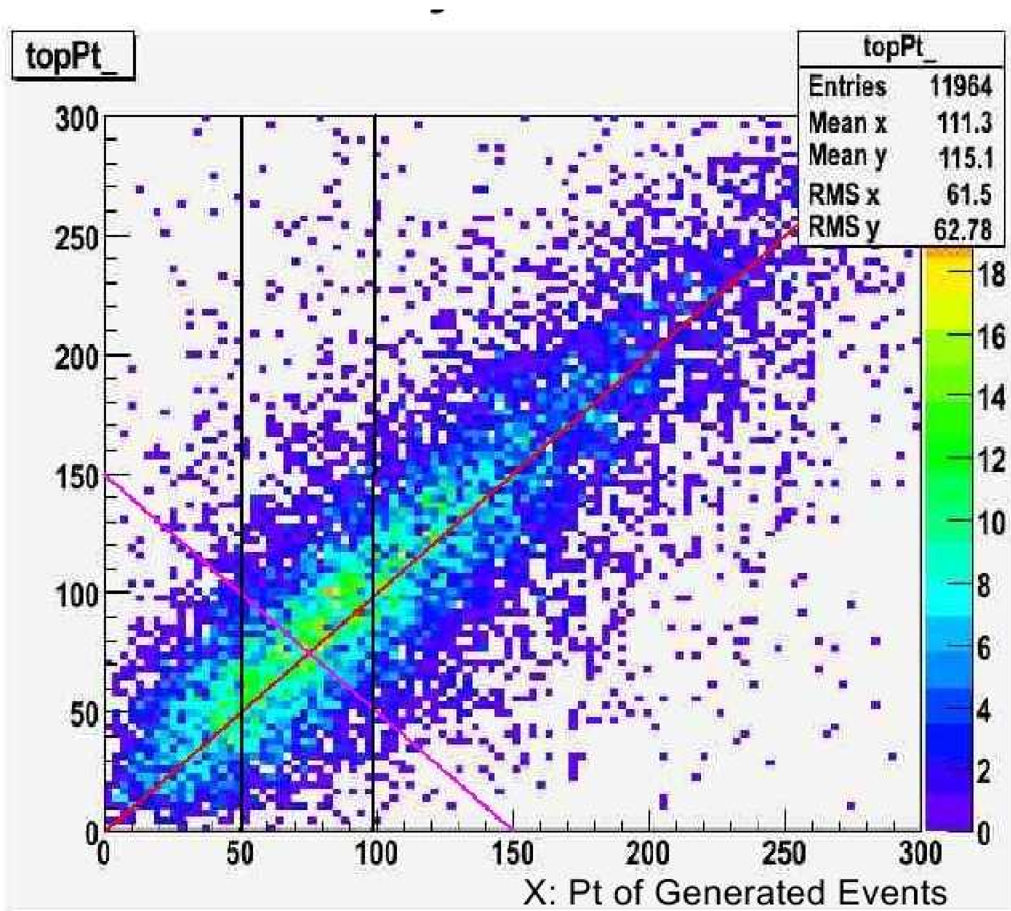


Figure 11: The correlation plot

The values are clearly correlated the correlation coefficient is very high Coff.=0,87. So for studying the distribution of the migration a projection of a cut perpendicular to the main diagonal is plotted for each bin. Than a Gaussian is fitted through the plot. The standard deviation is calculated from this plot When the bin size covers one std. in each direction than the stability and purity are expected to be about 0.6. In reality they are usually below this value. The world is unfortunately not perfect.

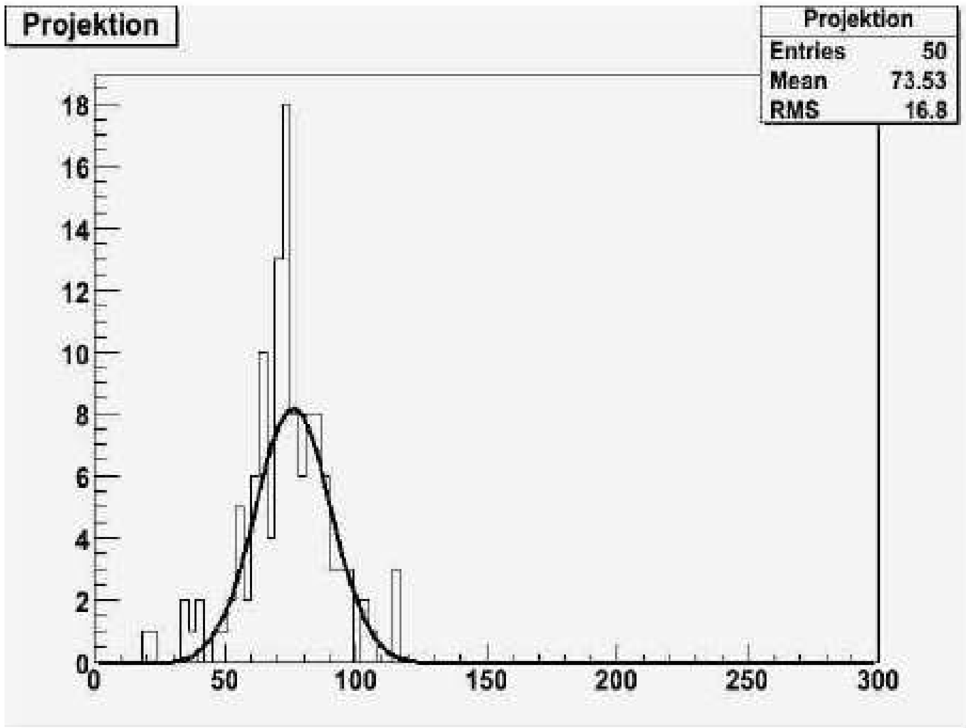


Figure 12: The cut through the correlation plot

As example the so found binning for the top Pt distribution is shown below.

Sqrt (2) * σ / GeV	Bin size / GeV	Bin setting / GeV
-	*10*	*0-10*
18,6	**25**	**10 – 35**
24,6	**30**	**35 – 65**
17,9	**35**	**65 – 100**
26,5	**35**	**100 – 135**
43,8	**35**	**135-170**
60,3	**45**	**170-215**
not significant	*85*	*215-300*

Figure 13: Example of a good binning

The bins chosen fulfill the requirement and a high stability and purity are expected. With the same method introduced above the binnings for all distributions were optimized as far the correlation allowed it.

3.2.1 The top kinematic variables

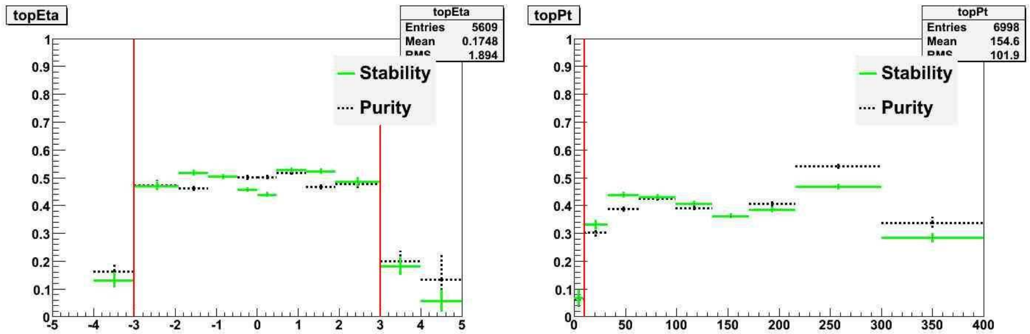


Figure 14: The stability and purity for the top quark

As can be seen in this plots the stability and purity are about 50% and what is very satisfying also is that the distributions are flat. That indicates that the events which migrate from the bin or into it do it not systematically but following a Gaussian distribution.

3.2.2 The t t-bar kinematic variables

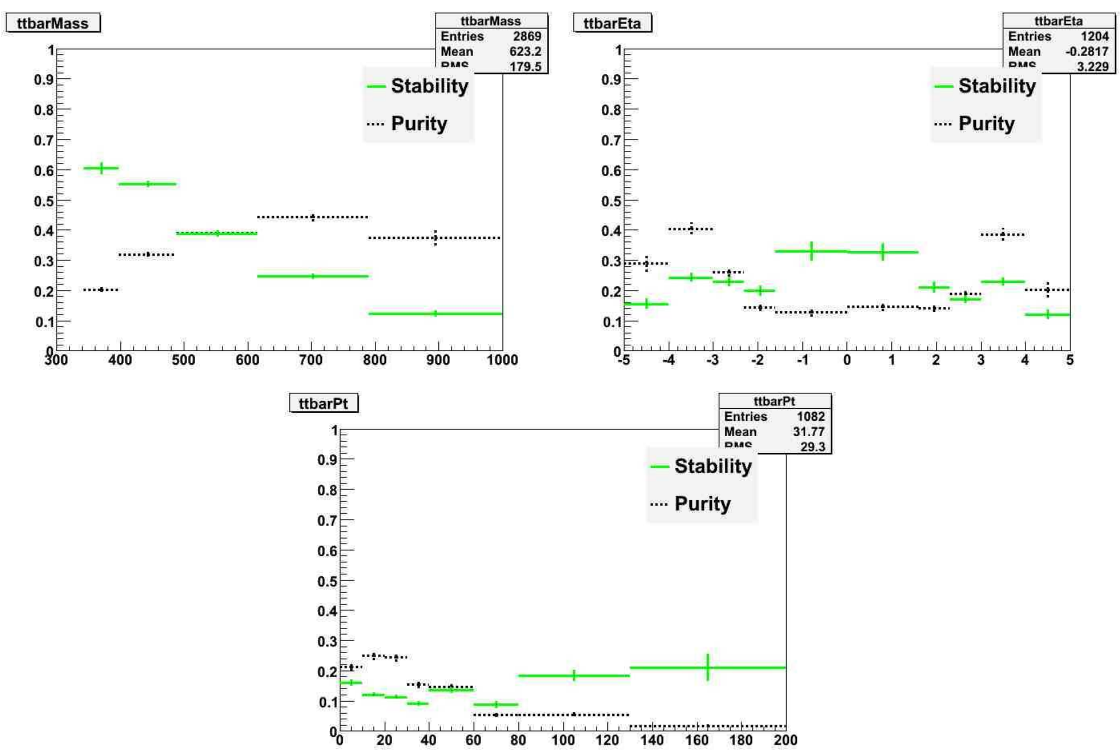


Figure 15: Stability and Purity for t t-ber kinematic vatiables

In the case of the t t-bar system the situation is less fortunate. In the Pt distribution systematic migration to lower Pt values can be observed. The other way round the mass of the t t-bar system is more frequently reconstructed with a higher value as it was generated. In the case of Eta there is systematic migration into the bins around the value of 4. Since in those bins the purity is high, that indicates that fewer events leave the bin, and the stability indicating the migration into the bin is rather low on the other hand. The outcome of the plots above is that the reconstruction of the t t-bar system needs to be done more carefully involving more sophisticated unfolding methods. This will be discussed in the fourth section.

3.2.3 The t t-bar Phi problem

In the two dimensional plot the correlation between the generated and the reconstructed Phi angels of the t t-bar system is shown.

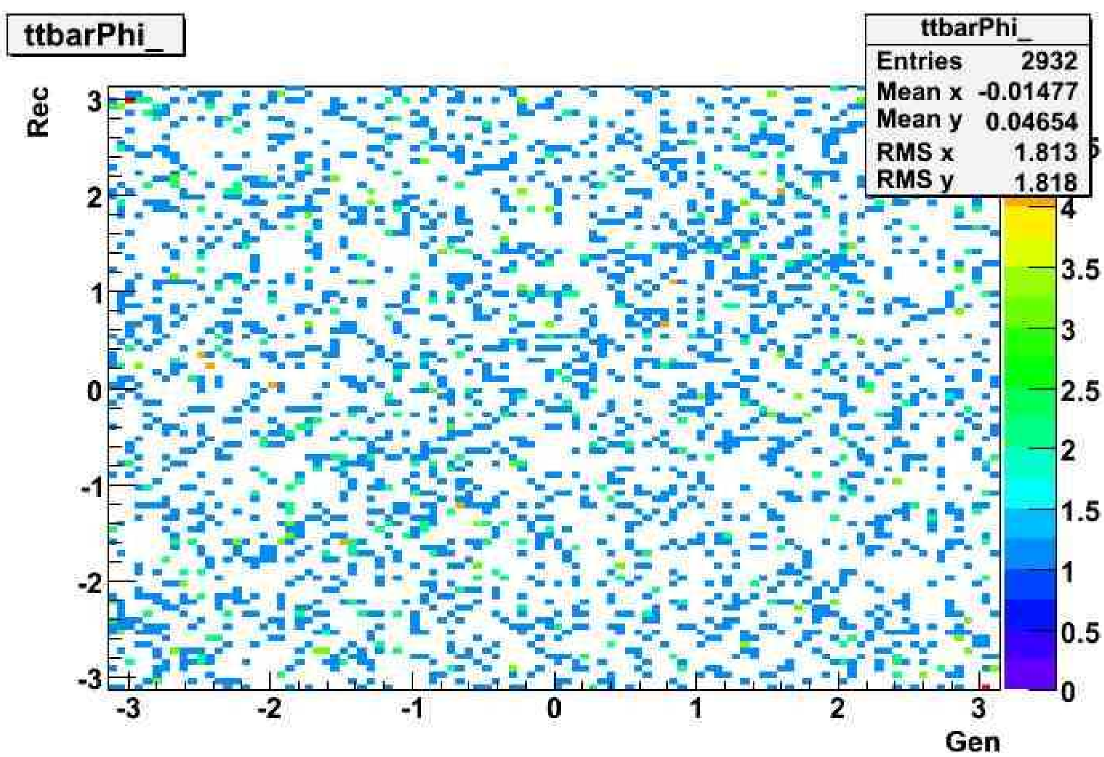


Figure 16: The t t-bar Phi correlation

As can be seen there is no correlation between the generated and the reconstructed events. This leak of correlation could not be properly explained. The wrong reconstructed neutrino might be the problem for the reconstruction of the whole t t-bar system. But this effect seen in the plot above is very severe and will have other reasons beside the neutrino as well. Concerning the total leak of correlation the Phi and Delta Phi values is not further regarded.

4 Further studies

4.1 The comparison of the hadronic/leptonic decayed top

As the first additional task the different behavior in the reconstruction of the leptonic and the hadronic decaying top quark was studied. In the Pt distributions of the hadronic and the leptonic top no significant differences could be found. But the Eta plots showed an interesting behavior. Below the left plot shows the hadronic decayed top and the right shows the leptonic decayed one.

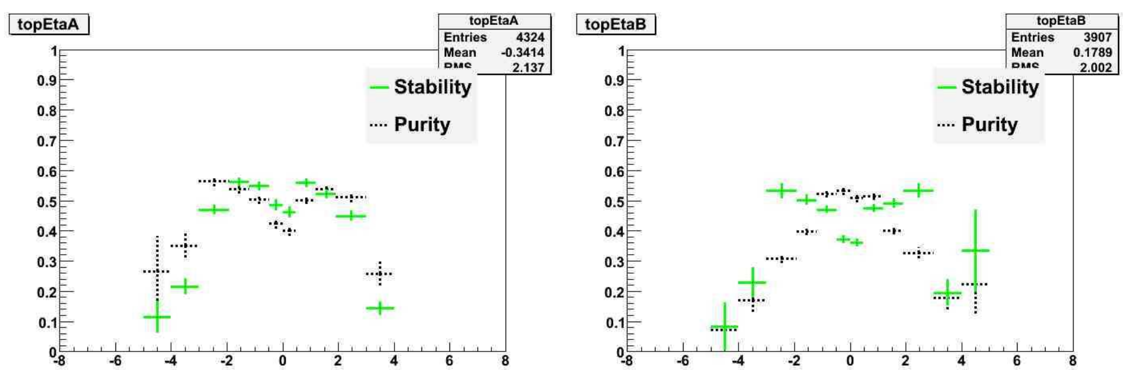
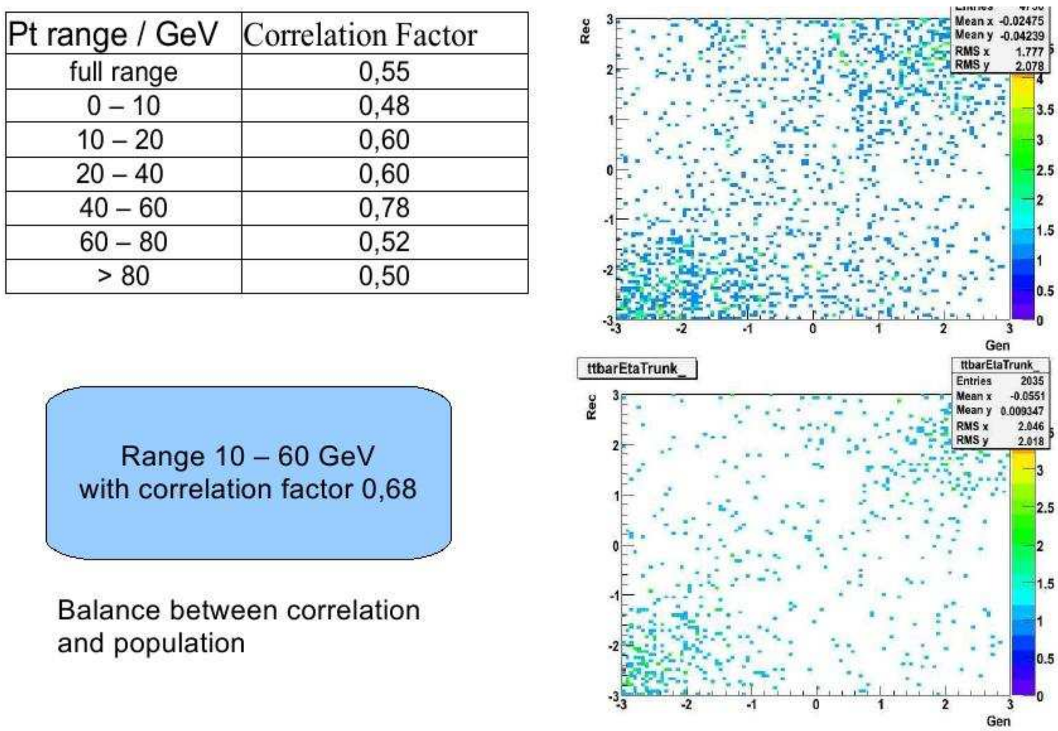


Figure 17: The hadronic and leptonic decay of the top

What can be seen here is that the stability and purity are of the same high in the bins on the hadronic side, while the stability around low Eta values goes down under the level of the purity on the leptonic side. This indicates migration into the Eta central region. One possible explanation is the neutrino on the leptonic side. The neutrino is reconstructed as missing transversal Energy and hence its z component is a priori zero. This causes a bias towards a lower value of the whole reconstructed quark. To avoid this effect an algorithm which calculates indirectly the neutrino's z component can be should be inserted into the CMSSW analysis framework.

4.2 The improvement of the t t-bar Eta resolution

Taking into account the correlation plot of the t t-bar Eta distribution it occurs an obvious correlation in the high Eta regions and rather uncorrelated data in the low Eta region. The idea now was to apply a cut on the Pt value of the t t-bar system to see whether this can improve the correlation.



Pt range / GeV	Correlation Factor
full range	0,55
0 – 10	0,48
10 – 20	0,60
20 – 40	0,60
40 – 60	0,78
60 – 80	0,52
> 80	0,50

Figure 18: The correlation Improvement of t t-bar Eta

Indeed the correlation can be improved when a Pt cut is applied. In the lower plot there is far less noise in the central region. On the generator level the optimal cut appears to be Pt between 10 and 60 GeV. This cut can improve the correlation factor from 0,55 to 0,68.

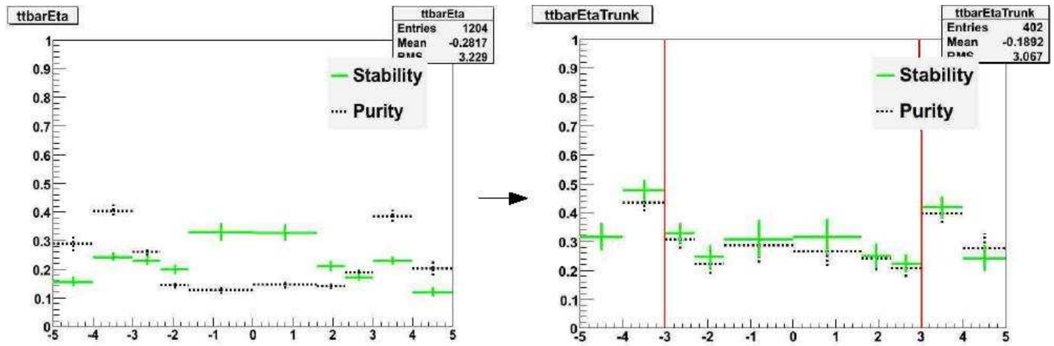


Figure 19: The improvement of the t t-bar Eta stability and purity

In the region of |Eta|<3 ,which is the visible region of the detector, stability and purity can be brought on a similar level and their distribution is flat and above 30%.

Of course cutting on the generator level is like playing God and pretending the absolute knowledge about the event. So in further studies Pt regions should be found which are applicable to the realistic situation.

4.3 The application of a realistic event hypothesis

4.3.1 The stability and purity of the top quark with the kGeom algorithm

Here the event hypothesis is constructed by applying a simple geometrical algorithm calls kGeom. Here geometrical properties like the cone of the jet and so on are used to match the jets to the particles which supposedly were participating in the reaction. In contrast to the kGenMatch algorithm -applicable to Monte Carlo data only- this algorithm can be applied to real experemental data as well.

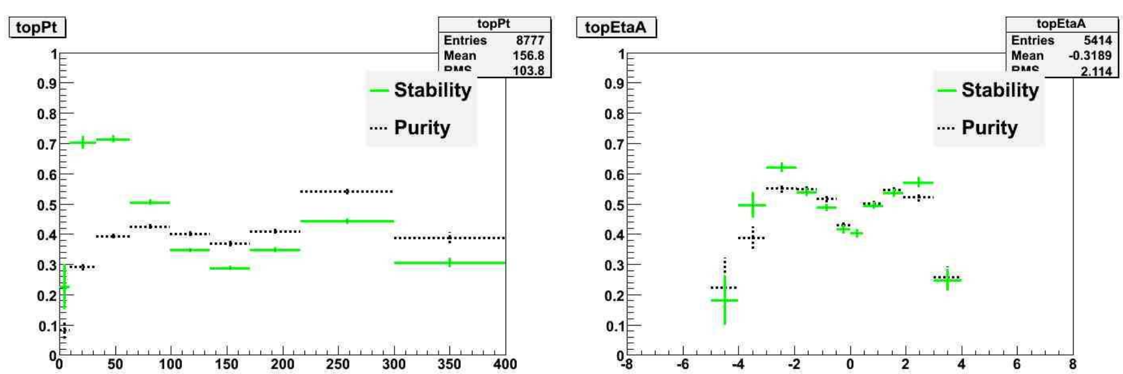


Figure 20: The Stability and Purity of the top applying kGeom

The Eta distribution above is from the hadronic decaying top only. For this top quark the stability and purity in the visible range of Eta remains on a high level between 40% and 60%. Also the biasing migration is tolerable. Still it will be interesting to study the system of both quarks after implementing the z-component reconstruction algorithm for the neutrino. The stability and purity for the Pt of the top quark is high as well. The distribution is flat above the Pt value of 50 GeV. This might be also an effect of this simple algorithm, that low Pt tops are reconstructed worse than those with a higher Pt value.

For the t t-bar system the stability and purity show strong systematic migration like in the case of the kGenMatch algorithm.

4.3.2 The improvement of the correlation of the t t-bar Eta applying the kGeom algorithm

While applying the kGeom algorithm and the cut on the reconstruction level it appeared as more useful to cut on the lower Pt only. Here a table summarizing those results:

Pt > x / GeV	Correlation Coef.
10	0,61
20	0,63
25	0,63
30	0,64
40 <Pt < 80	0,67
without Pt cut	0,57

Figure 21: Correlaton after Pt cut

So while without the cut the correlation factor was about 0,57 an improvement can be achieved. This correlation studies need to be done for every reconstruction algorithm which will be applied to estimate the optimal cut conditions.

4.4 Suggestions for the t t-bar Pt reconstruction

Since in this case a biasing migration appears it might be the best option to apply a detailed unfolding technique for the Pt of the t t-bar system. One possibility is to apply the χ^2minimizing technique: When y is the underlying distribution of the Pt and W[y] is the covariance matrix of the Gen and Rec distributions for the Pt. Than while $\chi^2$is minimized, one gains A as the unfolding matrix for the deconvolution.

$$\chi^2 = (Ax - y) \cdot W[y] \cdot (Ax - y)^T$$

This matrix can be used to obtain the real underlying distribution of the t t-bar Pt by the convolution of the W[y] with the (Ax-y) function.

$$x = \int W[y] \cdot (Ax - y)^T dy$$

After the successful reconstruction of the Pt distribution the Pt cut for the improvement of the Eta distribution should be show a similar good effect as the cut on the generator level.

5 Conclusion

For the kinematic variables of the top quark an appropriate binning could be found and the stability and purity of the distributions allow bin to bin corrections using the reciprocal of the efficiency in each bin. These results could also be gained using the realistic event hypothesis. So the technique is ripe to be applied to real data. The leptonic decayed top though needs more attention, but it is highly probable that its reconstruction can be highly improved by implementing an algorithm which determines the z-component of the outgoing neutrino. This will be done within the next months.

The number of bins of 6 to 8 in the top kinematic variables and their stable properties are suited very well to calculate the differential crossection of the top quark in grate detail. And this is the goal that future work should aim for.

Way more problematic is the reconstruction of the t t-bar system. Even if the proper neutrino reconstruction might improve those variables as well, more sophisticated unfolding techniques will be necessary for the reconstruciton. A strategy for instance could be to apply the 4.4 χ^2unfolding method to the Pt distribution and than to use those corrected Pt values to improve the Eta correlation by cutting out the proper Pt range. After that bin to bin corrections in Eta become applicable. The study of the t t-bar system is especially interesting when analyzing the gluon density in the colliding proton via the Eta boost. And the Pt of the t t-bar system is a value that could tell more about second order process of the top decay. So both systems the top and the t t-bar are worth studying but especially for the reconstuction of the t t-bar system there is still a lot of work to be done.

6 Acknowledgments

I would like to thank to:

- Dr. Roger Wolf, my supervisor in top group of the Hamburg University
- Sharham Aryan who was highly involved in this summer project
- Dr. Sebastian Naumann who helped me with the development of many new ideas for the project
- Niklas Pitsch, Holger Enderle and Matthias Schreoder who were always open for discussions and hat a big stock of coffee pads

References

[1] PDG Particle Physics Booklet *Particle Data Group* 2008