CHAPTER 5

DATA-DRIVEN QCD BACKGROUND DETERMINATION METHODS

In the previous chapter we mainly focused on the understanding of extra muons in $t\bar{t}$ like topologies. We have seen that most of these muons are non-prompt muons originating from semi-leptonic b-decay. The study was done on Monte Carlo, using truth information. In this chapter we will focus on data-driven methods to estimate the amount of QCD events that pass the $t\bar{t}$ cuts and especially the isolated muon cut, see Chapter 3. We explore two different methods from literature to achieve this goal [126]. This study has been performed on the $\sqrt{s} = 10$ TeV MC samples and serves as background to the method we will develop in Chapter 7 for the $\sqrt{s} = 7$ TeV data taken in 2010.

In Section 5.1 we present the strategy to do a data-driven estimation. Hereafter, in Section 5.2, we will show the input distributions that we use. Sections 5.3 and 5.4 will explain in detail the two methods. The results will be summarized in Section 5.5.

5.1 Strategy

We present two data-driven methods to estimate the QCD content after all the $t\bar{t}$ selection cuts [126]. Note that we will use selections close to the base selection presented in Section 3.3.1 and not the two extra requirements presented there (Sections 3.3.2 and 3.3.3), nor will we use the overlap removal of muons that are close to jets. The reason for this is that we want to make use of distributions of non-isolated muons originating from QCD to extrapolate into the isolated signal region. The overlap removal of muons to jets works like an isolation cut and leaves us with too little statistics in the non-isolated region. The first method is called the *ABCD method* and makes use of two uncorrelated observables in four regions of their two-dimensional phase space. We construct the regions such that only one of the regions is dominated by signal and the information of the other three is used to estimate the background contribution in the signal region. The second method, the *fit method*, is based on the observation that the

isolation of muon tracks is very different for signal muons compared to QCD muons. This information is used to fit the isolation distribution in a QCD dominated region and extrapolate into the signal region. Both methods rely on the fact that there is an observable that behaves differently for signal and background and that one can find a region for that observable that is dominated by background.

5.2 Input distributions

In order to separate signal and QCD-background we need observables that differ between the two. In Chapter 4 we have seen and studied the distributions of some of these observables. The main differences were seen in the isolation of the muons and their distance to the closest jets. However, in order to use the ABCD method in its simplest form (described further in Section 5.3) the distributions used have to be uncorrelated. This is not the case for the isolation of a muon and its distance to the closest jet. Instead we exploit another variable which represents our knowledge that the extra muons mainly come from semi-leptonic b-decays: the impact parameter significance. We investigate both the muon isolation and the muon impact parameter significance in more detail.

5.2.1 Absolute isolation and relative isolation

Until now we have used the absolute isolation of a muon. The two methods to estimate the QCD contribution in a data-driven way that we will explore in this chapter, have however been developed using a relative isolation: $E_T^{dR=0.20}/p_T$ [1]. In Figure 5.1 the relative isolations is shown for prompt muons in $t\bar{t}(\mu)$ and for the muons in the QCD sample.
We note that the distribution resembles the one that was shown in the previous chapter, Figure 4.19 (right). To compare the two, the fraction of muons that passes a given cut is shown in Table 5.1. By placing the cut at 0.14 the efficiency for prompt muons is the same as it was for the absolute isolation cut, but with higher rejection for QCD muons. The cut for the relative isolation variable has been put at 0.10. It can be seen that cutting at 0.10 on the relative isolation yields a little lower signal efficiency, but also a much higher background rejection.

	rejection of muons	
	prompt muons	extra muons
$E_T^{dR=0.20} < 6$ GeV	4%	67%
$E_T^{dR=0.20}/p_T < 0.14$	4%	83%
$E_T^{dR=0.20}/p_T < 0.1$	6%	90%

Table 5.1: *Percentage of prompt ($t\bar{t}(\mu)$) and extra (QCD) muons that does not pass a given isolation cut.*

[1] $E_T^{dR=0.20}$ is the transverse energy in a cone of dR = 0.20 around the muon.

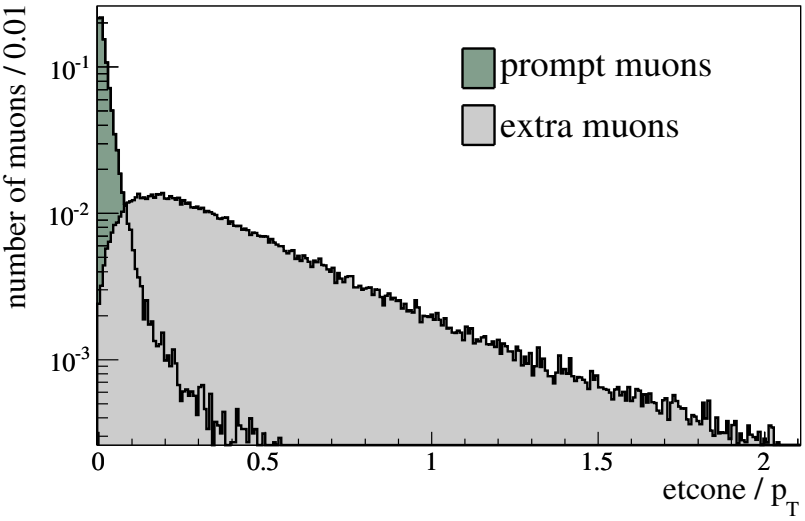

Figure 5.1: *The $E_T^{dR=0.20}/p_T$ distribution after all other muon cuts for prompt muons in $t\bar{t}(\mu)$ (dark) and extra muons in QCD (light), normalized.*

The isolation variable is clearly a good variable to distinguish signal and background. The reason behind this is the difference in production mechanism for prompt and extra muons, as explained in Chapter 4.

5.2.2 Impact parameter significance

We have shown in the previous chapter that most of the extra muons originate from semi-leptonic b-decay. There is one feature of this type of decay that hasn't been exploited yet: B-mesons live relatively long. The average lifetime of bottom-mesons is of the order of 10^{-12} s, which means they travel on average a few mm before they decay (for a b-jet p_T typical for the $t\bar{t}$ topology of around 60 GeV). This feature is represented in the d0 variable, which is the distance of closest approach in the transverse plane between a track or object and the interaction point, see Figure 5.2 (left). The distance in the transverse plane is used since the boost in the z-direction can be large due to large differences between x_1 and x_2 see Section 1.2. However, due to beam-spot width and displacement most interactions will not be located at the center of the ATLAS coordinate system, but at a primary vertex, that may be displaced from it. The d0 variable that is quoted the rest of this chapter is calculated with respect to this primary vertex (and is sometimes referred to as the corrected d0). In Figure 5.2 (right) the d0 of prompt ($t\bar{t}(\mu)$) and extra (QCD) muons is shown. The distributions are normalized to unity and we note that the QCD muons give rise to a broader distribution.
The error on the d0 of a muon can be quite large and therefore a more powerful discrimination between signal and background can be achieved by using the significance of the d0 parameter, which is defined by: d0 significance = d0/σ(d0). The uncertainty

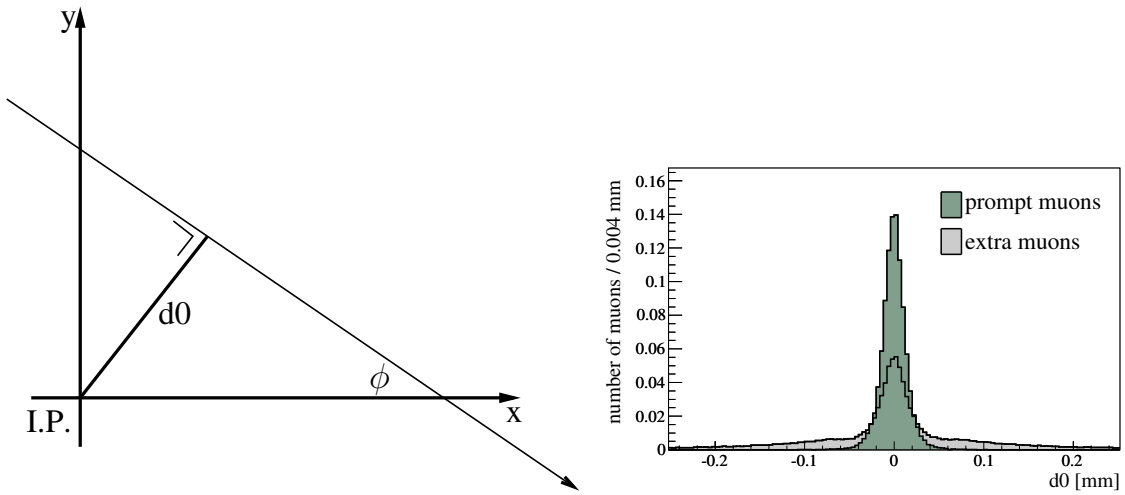

Figure 5.2: *Left: when a track traverses the detector, d0 is defined as the transversal distance to the interaction point (I.P.). Right: the d0 distribution for prompt muons in $t\bar{t}(\mu)$ (dark) and QCD muons (light), normalized to unity.*

on the d0 significance ($\sigma(d0)$) is given by the uncertainty that comes directly from the measurement of the d0, but also has to contain the uncertainty on the primary vertex which was used in the calculation of the corrected d0. The full $\sigma(d0)$ definition that we use is given by:

$$\sigma(d0) = \sqrt{\sigma_{d0}^2(\mu) + \sin^2(\phi)\sigma_x^2 + \cos^2(\phi)\sigma_y^2 - 2\sin(\phi)\cos(\phi)\sigma_x\sigma_y}, \tag{5.1}$$

where ϕ is the angle in the transverse plane of the muon track, $\sigma_{d0}(\mu)$ is the uncertainty on the d0 of the muons and σ_x (σ_y) the uncertainty on the x (y) parameter of the primary vertex. The $\sigma_{d0}(\mu)$ and the σ_x/σ_y are of the same order of magnitude, but due to the angle ϕ terms in Equation 5.1, the uncertainty on the d0 of the muon is the dominant source of the uncertainty of the $\sigma(d0)$ parameter. In Figure 5.3 the d0 significance of both prompt muons in $t\bar{t}(\mu)$ (dark) and QCD muons (light) is shown. We observe that prompt muons have a steeper peak at low values.

Both the impact parameter significance and the relative isolation are variables to distinguish between signal and QCD background. Note that the background that we consider here is only QCD. Muons from W+jets or other top channels will in these distributions end up in the signal region since the muons are in principle prompt muons (from a W) and cannot be distinguished from the muons in $t\bar{t}(\mu)$. Those backgrounds have to be determined with other methods, as was discussed in Section 3.3.

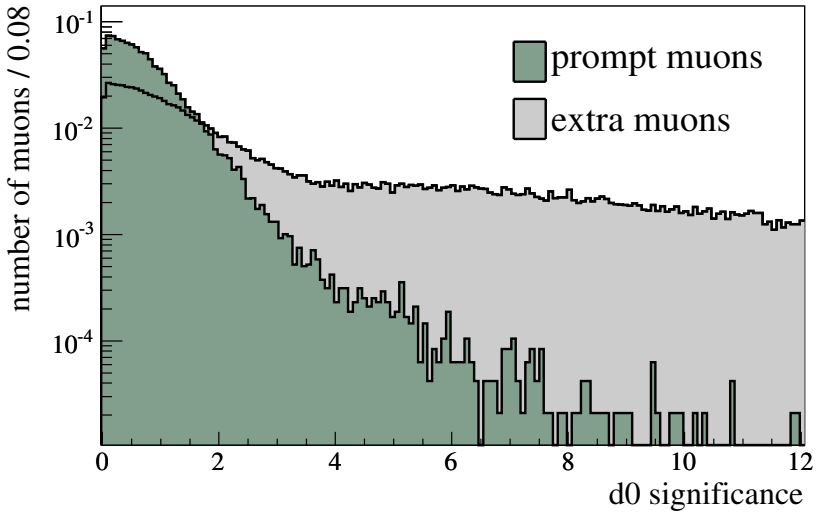

Figure 5.3: *Impact parameter significance (d0 significance) for prompt muons in $t\bar{t}(\mu)$ (dark) and QCD muons (light), normalized to unity.*

5.3 ABCD method

The ABCD method relies on the fact that one has two independent distributions to distinguish between signal and background. We have shown in the previous section that the impact parameter significance (d0 significance) and the relative isolation are powerful variables to distinguish between prompt and non-prompt muons. In Figure 5.4 the d0 significance versus relative isolation (after all other muon cuts and shown for events with 2 jets) is shown, where we divided the distribution in four boxes. The boundaries of these boxes are chosen such that C is the signal region or put differently: all other regions should be signal free. This is ensured also by not letting the boxes connect. The choice of the boundaries for the relative isolation is on the one side the normal cut-value of 0.1 and on the other side a larger value where no prompt muons should be found: 0.15, see Figure 5.1. For the d0 significance boundaries the same arguments hold: a lower one at 3 since almost all signal will be contained within, see Figure 5.3 and a higher one to have a signal-free region. Thus the four boxes are defined as:

A :	d0 significance < 3	relative isolation	$>$	0.15
B :	d0 significance > 5	relative isolation	$>$	0.15
C :	d0 significance < 3	relative isolation	$<$	0.10
D :	d0 significance > 5	relative isolation	$<$	0.10

If the two assumptions mentioned here hold (the distributions are independent and

regions A,B and D are signal muon free) then the ratio of the content of A and C should be equal to B and D (also A/B and C/D should be equal, the system is over constrained).

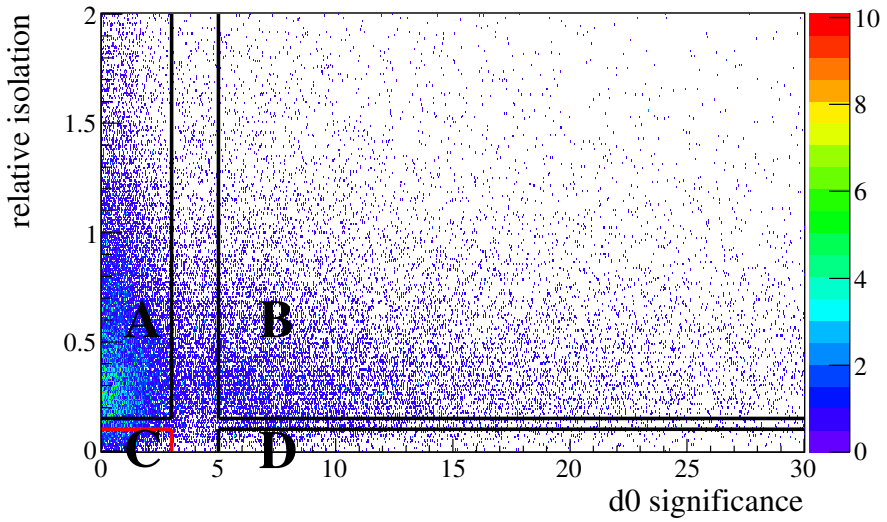

Figure 5.4: *The d0 significance versus the relative isolation distribution after all other muon cuts for muons from $t\bar{t}$, QCD, W+jets and single top in 2 jets events.*

The above explanation leads to the simple expression:

$$\mathrm{C} = \mathrm{A} \cdot \frac{\mathrm{D}}{\mathrm{B}}, \tag{5.2}$$

where C is the number of QCD events in the region where we expect $t\bar{t}$ to end up. We can thus calculate the number of QCD events that will pass all our $t\bar{t}$ selections and will be mistaken for signal.

5.3.1 Independent distributions

We can check that these distributions are indeed independent by looking at the correlation coefficient that is shown for different jet multiplicities in Table 5.2. The correlation coefficient (ρ) of two variables x and y is defined as[2]: $\rho = \frac{<xy>-<x><y>}{\sigma_x \sigma_y}$.
The correlation coefficient lies in the interval [-1,1], where 1 (-1) means that the variables are fully (anti-)correlated. A ρ of zero means that x and y are uncorrelated. As can be seen in the table, the relative isolation and d0 significance are hardly correlated except in two jet events. This means that the assumption of independent variables only holds truly for higher jet multiplicities. Note that the correlation coefficient can be calculated from data and the assumption can thus be tested. The independence of the variables

[2]The error on ρ when N is large is given by: $\sigma_\rho = \frac{(1-\rho^2)}{\sqrt{N-1}}$

number jets	ρ
all	0.103 ± 0.003
2	0.139 ± 0.004
3	0.084 ± 0.005
4	-0.002 ± 0.005

Table 5.2: *Correlation coefficient of the relative isolation and d0 significance variable of muons in the QCD sample for events with 2,3 and 4 jets.*

means that the distribution of relative isolation looks the same at any given value of d0 significance (the same is true the other way round). That this is the case can be visualized by plotting the relative isolation in slices of the d0 significance and by looking at d0 significance in slices of relative isolation, see Figure 5.5. The total number of events per slice differs and the plots are normalized to show that they have indeed the same shape.

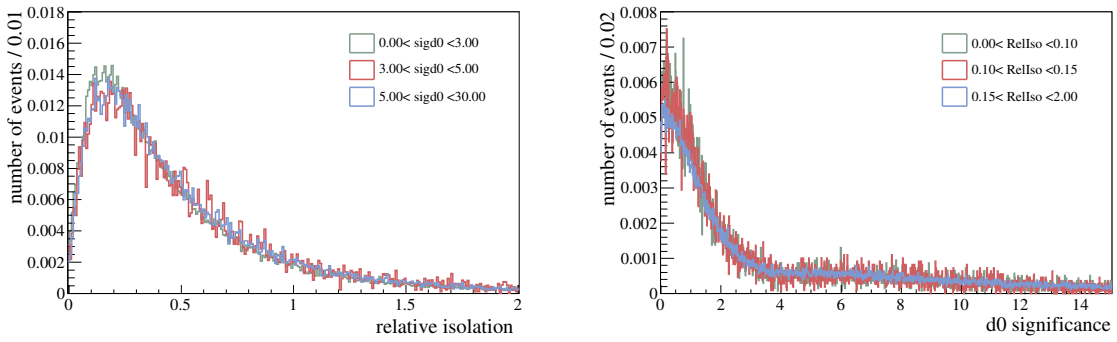

Figure 5.5: *Left: relative isolation of QCD muons after all other cuts in slices of d0 significance normalized to unity. Right: d0 significance of QCD muons after all other cuts in slices of relative isolation normalized to unity.*

This feature is the reasoning behind the method, since if the distribution of one variable is the same for different slices of the other variable then the ratios that we mentioned above should indeed be equal.

5.3.2 Application of the method

If we now count the events from QCD, the other major background samples and the signal samples, we get the results presented in Table 5.3. Here we show the number of events in each region for QCD, signal ($t\bar{t}$ (μ)), W+jets together with single top and total number of events. In the left table this is presented in the case of two jet events, hence the contribution of non-QCD events is low. In the right table the numbers are presented for events that pass all top cuts and we see that region C is now dominated

by prompt muons. Note here that the regions A and B are virtually signal free. The region D however contains almost as many signal-like muons (i.e. from $t\bar{t}$, W+jets or single top) as QCD muons. Most of these muons can be traced back to τ decays in $t\bar{t}(\tau)$ or $W(\tau)$ events. This obviously violates the second assumption.

2 jet events					4 jet events + all other top cuts				
	QCD	$t\bar{t}$ (μ)	W+jets, single top	total		QCD	$t\bar{t}$ (μ)	W+jets, single top	total
A	22263.9	1.3	49.7	22318.0	A	1345.7	12.9	8.7	1392.3
B	16497.1	0.7	4.4	16504.1	B	1375.5	6.3	3.2	1416.7
C	1531.3	31.0	2909.5	4504.2	C	12.7	153.9	119.5	311.6
D	982.2	0.1	38.8	1021.7	D	6.5	0.7	2.4	11.9

Table 5.3: *Number of events from different samples per region for 9.6 pb^{-1}. Left: in events with 2 jets. Right: in events that pass all top event selection cuts. Note that $t\bar{t}(\mu)$ is shown separately, but the total includes all other $t\bar{t}$ events.*

If we use the expression for C to predict the number of QCD events in the signal region we get the results show in Table 5.4, where we show the results for various jet multiplicities and on the last row for events after all top cuts.

Jets	QCD predicted	QCD MC
2	1381.6 ± 44.7	1531.3 ± 44.4
3	437.2 ± 23.2	501 ± 22.4
≥4	199.6 ± 15.2	217.4 ± 14.7
all top cuts	11.7 ± 2.6	12.7 ± 3.6

Table 5.4: *Predicted and MC number of QCD events in region C for 9.6 pb^{-1}. First three rows without $\not{E}_T$ constraint, last row with all top cuts.*

From Table 5.4 it is clear that the predictions, although too low for the 2 jet case, are very good for the higher jet multiplicity events. The quoted error is only the statistical error and clearly does not reflect the full uncertainty on the prediction. Note that the statistical error here is not $\sqrt{N}$, with N the number of events. This is due to the fact that all samples are scaled down to represent 9.6 pb^{-1} (as the QCD sample) and also the negative weights of `MC@NLO` are taken into account. In order to find the systematic uncertainty that is associated with this method we varied the boundaries of the method. By shifting either the upper bound of the d0 significance or the upper bound of the relative isolation we left the signal region C untouched. By varying these boundaries we evaluated the change in the predicted number of events compared to the baseline prediction. The results are shown in Figure 5.6 where we show the shifted prediction divided by the baseline value. This by construction gives value 1 for the d0 significance at 5 and for the relative isolation at 0.15.

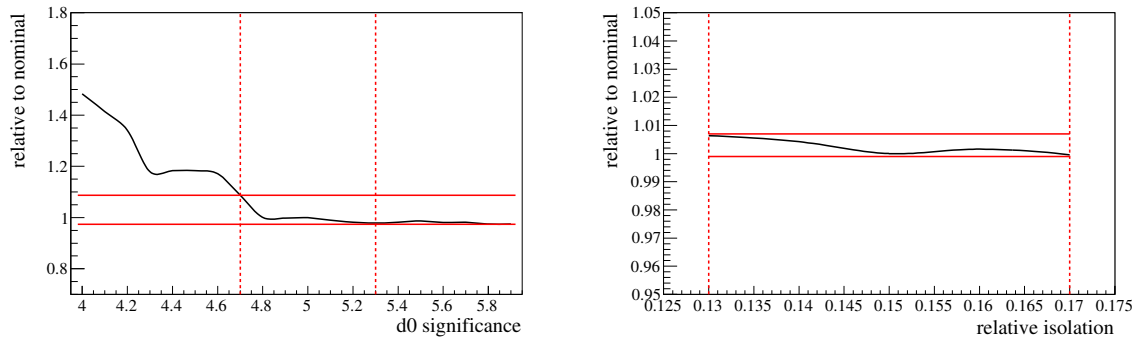

Figure 5.6: *Relative change of the predicted number of QCD events in region C due to change of boundaries. Left: changing d0 significance. The red lines denote what has been taken as systematic uncertainty. Right: changing relative isolation.*

The first thing that has to be noted are the apparent jumps in the plot (left) of the d0 significance. This has been investigated and traced back to single QCD events with high weight migrating from one region into another. This effect is thus due to limited statistics of the sample. The rise however that can be seen is caused by the isolated muons (from signal or other non-QCD background) entering region D. When moving to lower values of the upper d0 significance boundary this is likely to happen more frequently. The nominal value of 5 is on a plateau and thus a good choice. We assigned a systematic uncertainty of +9% and -3% to this systematic corresponding to the change in the prediction when varying the value of d0 significance ±0.3. On the right side, the relative isolation, we note that the prediction is stable with respect to changing the boundary. The change is smaller than 1%. The overall result of this method is then:

$$\mathrm{C} = 11.7 \pm 2.6\ (\mathrm{stat})\ ^{+1.1}_{-0.4}\ (\mathrm{syst}) \tag{5.3}$$

5.3.3 Conclusion

The ABCD method in its most simple form relies on the use of two independent distributions that can distinguish between prompt and non-prompt muons. We have shown that we can predict the number of QCD events in the signal region with the method. One of the regions however is not prompt muon free (as we assume) and this should ideally be taken into account. Since the system is over constrained one option would be to introduce an efficiency for prompt and non-prompt muons to enter D. This system is then still solvable and accounts for the contamination. One apparent problem of the method at this stage is that we end up calculating the amount of QCD in a region that is not exactly the same as the nominal $t\bar{t}$ selection region. We introduced a new cut: the d0 significance cut. There are good reasons however to use this cut anyway. First of all it removes almost no signal (154 compared to 157 events after all cuts in 9.6 pb^{-1}). Secondly it not only removes QCD muon, it also protects the analysis from cosmic muons since they do not originate from the interaction point. By studying the

systematic error associated with varying the boundaries it became clear that shifting towards lower values of the d0 significance parameter the estimation blows up. Since this however can be measured from data and one can put the higher bound at a value where the distribution levels off, see Figure 5.6, this should not be a problem. We will see in Chapter 6 however that the statistics in the early data is too low to use this method, but we will develop a method that is closely related to the ABCD method.

5.4 Fit method: extrapolation of the isolation variable

The second method that we will use to predict the number of QCD events that pass the top event selection uses the shape of relative isolation for muons in QCD events. By measuring the distribution of relative isolation of muons after all other muon cuts and fitting it in a region that contains no signal-like muons (no isolated muons from $t\bar{t}$ or W+jets), we can extrapolate the distribution into the signal region. The integral of this extrapolation below the cut value then estimates the number of QCD events in the top selection. In Figure 5.7 the distribution of relative isolation is shown of all muons that pass all cuts except the isolation cut in two jet events (left) and events that pass all top event selection cuts (right). The samples used here are the QCD samples, W+jets, single top and the full $t\bar{t}$ sample. The fit is performed from 0.1 to 2 and then extrapolated down to the signal region (dark line). For comparison we also show the actual QCD spectrum.

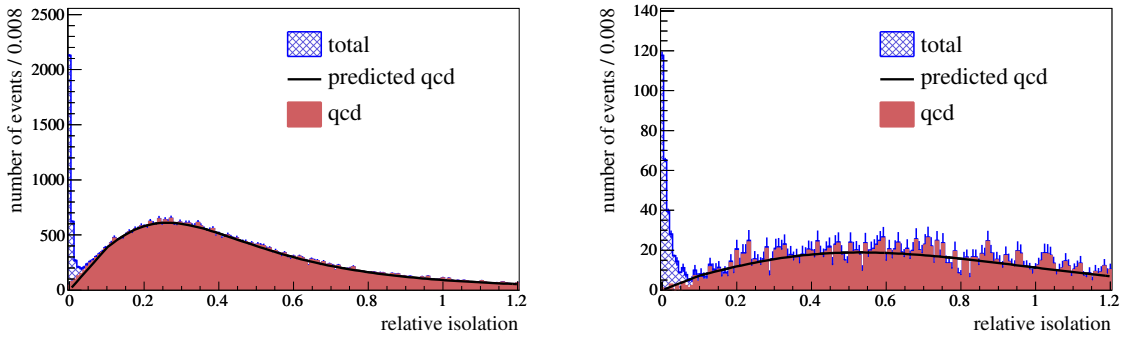

Figure 5.7: *Relative isolation of muons after all other muon cuts for $t\bar{t}$, W+jets, single top and QCD (hashed) normalized to 9.6 pb^{-1}. The black line is the prediction (plus the fit) and the light filled graph the QCD spectrum. Left: in two jet events. Right: events that pass all top-cuts.*

The fit that we use is a Landau-function multiplied by a third order polynomial and the extrapolation uses a straight line from (0,0) to the end-point of the fit. This linear extrapolation was chosen in order to not be too dependent on the exact shape of the isolation distribution in the signal region. We noticed that the shape in Monte Carlo

changed when going to higher jet multiplicities between being convex and concave and thus the straight line is a simple and stable average. We can see that indeed above 0.1 there is hardly any signal ($t\bar{t}(\mu)$) or other signal-like isolated muons (the difference between QCD and the total above 0.1 is negligible). The extrapolation yields the numbers as given in Table 5.5, where we show the integral (I) of the extrapolated fit (relative isolation < 0.1) compared to the MC number of QCD events in the region.

Jets	QCD predicted	QCD MC
2	2403.4 ± 203.3	3112.6 ± 55.8
3	852.5 ± 77.5	1057.5 ± 32.9
≥4	468.8 ± 72.1	485.3 ± 22.9
all top cuts	45.3 ± 6.1	34.6 ± 6.2

Table 5.5: *Predicted and MC number of QCD events in signal region (relative isolation < 0.1) for 9.6 pb^{-1} in different jet multiplicity bins.*

The error that is quoted in Table 5.5 can be seen as the statistical error that comes from the fit. By varying the parameters within their uncertainty we get the plot that is shown in Figure 5.8, where we can see what each parameter variation does to the total fit. Since we are only interested in the error on the y-value of the endpoint for the extrapolation, we used the following formula for the error matrix of a function given the error matrix of the different variables [127]: $V(x) = GV(f)G^{-1}$, with G the vector with elements $\frac{\partial y}{\partial p_i}$ and $V(x)$ the covariance matrix. In this notation y is the function that determines the endpoint of the fit and p_i are the different parameters. More precise for this case the vector G is filled with the change of the value y of the endpoint due to a change of a given parameter within its error divided by that error: $\frac{\partial y}{\partial p_i} = \frac{\Delta(y(p_0) - y(p_i + \sigma_{p_i}))}{\sigma_{p_i}}$. We then used the error on the endpoint to calculate the error of the extrapolation.
The systematic error has been calculated for this method in the same manner as for the ABCD method: by varying the boundary. In this case we have varied the starting point of the fit (nominal at 0.1) and we found that the fit is stable around the nominal starting value (in a range of 0.7 - 1.3) and we have assigned a systematical error of ±10%. This summarizes the fit result as follows:

$$\mathrm{I} = 45.3 \pm 6.1\ (\mathrm{stat}) \pm 4.5\ (\mathrm{syst}) \tag{5.4}$$

5.4.1 Conclusion

The fit method relies on the fact that the distribution of the relative isolation of muons after all other muon cuts above 0.1 is dominated by QCD muons. This distribution can then be fitted and extrapolated into the signal region. We have shown that with a linear extrapolation we can predict the number of QCD events in the signal region with reasonable errors. The method is based on the idea that we can measure this relative isolation distribution in data and that we can also clearly see the kink that marks the

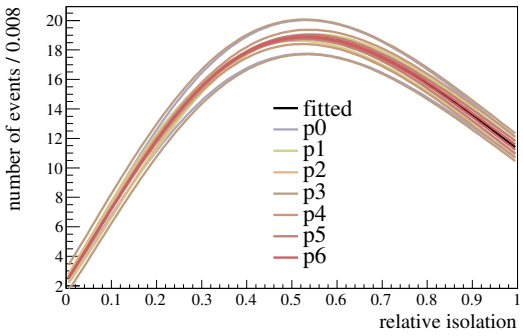

Figure 5.8: *Variation of the fit when varying the parameters within their uncertainty, of importance for us is the value at 0.1 (the endpoint).*

transition between prompt and non-prompt (mainly QCD) muons. If this kink is not there then one has to be careful with the assumption that there are no prompt muons above 0.1. A way to check this could be to try to find regions that are enriched with QCD (like 2 jet events) or regions that are QCD free by only investigating muons after overlap removal to jets which should be very isolated.

5.5 Summary

In this chapter we have investigated two methods to determine the QCD content after applying the base selection cuts. Both the ABCD method and the fit method give good results with total errors (statistical and systematical) under 20% with only 9.6 pb^{-1}. The dominating errors are statistical. The main problem with the ABCD method as it is now is that it predicts the number of QCD events after one extra cut: the d0 significance. This is also the reason that both methods cannot be compared directly. In order to do so one would either have to cut on d0 significance also for the fit method or find a way of returning not only the number of events in region C, but also the QCD content of D and the region in between. On the other hand however, the d0 significance cut has more advantages, like rejecting cosmic muon, and might be worthwhile considering as standard cut.

This study was performed using the $\sqrt{s} = 10$ TeV samples. We know however that the first collisions will occur at a center of mass energy of 7 TeV. We will see that with the lower statistics in the data sample and with the tighter cuts that will be used for the first top-quark pair production cross section measurement (including the overlap removal of muons close to jets) that the two methods presented here will not be adequate, see next chapter. In Chapter 7 we will develop a new method that is related to the here presented ABCD method using the knowledge that was gained in the last two chapters.