CDF/DOC/TOP/CDFR/5239
February 26th, 2000

On the Significance of the Dimuon Mass Bump and the Greedy Bump Bias

Tommaso Dorigo, Michael Schmitt

Harvard University

Abstract

The dimuon mass peak described in CDF-4860 is apparently a compelling reason to believe in the presence of a scalar quark signal in the CDF data. We investigate here the problem of determining the true significance of that signal with a toy Monte Carlo technique. Given the importance of that number, we defer a precise determination of the probability to a future version of this note, using the method described here, when an agreement will have been found with the authors and godparents on some details of the computation.

While analysing this problem, the authors have come across a very interesting, and general, bias suffered by likelihood or chisquared fits to a spectrum containing a large background and a small signal whose mass is a priori unknown. The bias is discussed and explained.

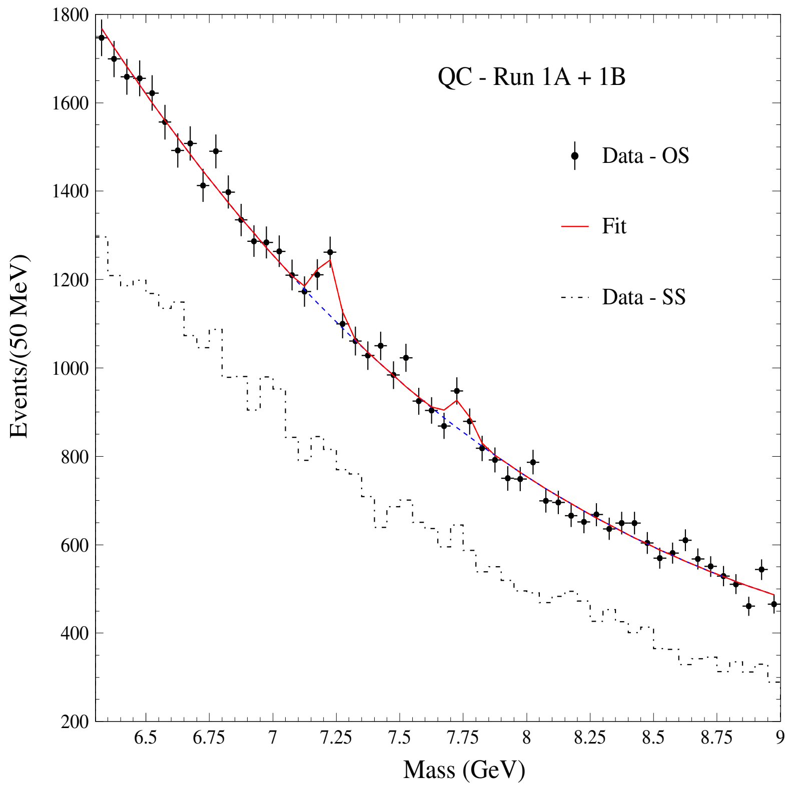

Figure 1: *The fit to the dimuon spectrum (from CDF-4860).*

1 Introduction

In CDF-4860 a signal for a bound state of two scalar bottom quarks is sought in the whole Run-1 dimuon dataset[1]. The authors of that analysis move from the observation that a scalar bottom quark of mass between 3.2 and 4 GeV provides a satisfactory explanation to the kinematic characteristics of an otherwise puzzling excess of events containing two soft leptons, extracted from a detailed comparison of the inclusive lepton data with Monte Carlo simulations subjected to a careful tuning of their heavy flavor content[2].

Such scalar quarks would have no reason to avoid making a bound state; this would be a quarkonium-like particle with a cross section roughly equal to or larger than 2.3% of the Upsilon(1S) resonance[3]; assuming that its branching ratio to a pair of muons is equal to

[1]G. Apollinari, I. Fiori, P. Giromini, F. Happacher, S. Miscetti, F. Ptohos, *Measurement of the dimuon invariant mass spectrum between the ψ and the Υ*, June 10th, 1999, CDF/ANAL/TOP/CDFR/4860_1a1b.

[2]See G. Apollinari, I. Fiori, P. Giromini, F. Happacher, S. Miscetti, F. Ptohos, *Observation of semileptonic decays of heavy quarks produced in $p\bar{p}$ interactions and the existence of one scalar quark*, CDF/ANAL/TOP/CDFR/4795, Feb. 6th, 2000, p. 42.

[3]G. Apollinari *et al.*, "Measurement"...

that of the Upsilon, about 250 such decays would find their way to our dimuon datasets. An optimization of the cuts necessary to clean the Upsilon signal produces 53,242 events in the dimuon mass region between 6.3 and 9.0 GeV, where the spectrum resembles a falling exponential, far from either trigger turnon effects and Upsilon contaminations.

The authors fit that spectrum with a binned chi-square fit to the sum of three gaussian shapes, the first of which models the exponentially falling background, and the latter two model a possible onium signal and even an excited state. The results are shown in Fig. 1. From now on we will disregard the excited signal, whose evidence is clearly not compelling, and we will concentrate on the larger bump sitting at a mass of 7.2 GeV.

The fit returns a total of 249.7 ± 60.9 events when a signal width of 38 MeV is fixed. Other fits performed with unbinned likelihood techniques seem to approximately confirm that result, but they prefer a smaller width if that parameter is allowed to float –and consequently a smaller number of events. We will not discuss in the following the problem of establishing whether the data are compatible or not with a 38 MeV width. Back-of-the-envelope calculations based on the returned chisquared of fits with different binning and widths seem to confirm that a width of 38 MeV is a fair interpretation of the data[4], and –although a deeper investigation would be nice– the issue is neglected here for the time being.

The significance of the bump at 7.2 GeV is a number of great importance because it seriously conditions our interpretation of the anomalies found in the $W + jets$ and the inclusive lepton datasets[5], and henceforth our decisions concerning the publication process. We will estimate that significance number here using a toy Monte Carlo approach. The technique is straightforward but its handling is not, since assumptions have to be thrown in if a meaningful result is to be extracted. We will discuss these issues in the following.

2 The Toy Monte Carlo Technique

A pseudoexperiment can be generated from a background shape similar to what is found in the data by just drawing randomly 53,242 events from it. No signal is present in the resulting spectrum, but Poisson fluctuations will at times resemble a gaussian bump. We can fit this spectrum with an unbinned likelihood technique to a function equal to the generating background shape plus a gaussian with a fixed 38 MeV width. The mass is left allowed to float in the fit, but an input value needs to be specified. Let's forget about this input value for a moment.

[4] B. Carithers, personal communication

[5] See G. Apollinari, M. Barone, I. Fiori, P. Giromini, F. Happacher, S. Miscetti, A. Parri, F. Ptohos, *A study of the kinematics of the* $l+ \not{E}_T + 2,3$ *jet events with multitags confirms the evidence for non Standard Model physics with the statistical significance of a discovery*, CDF/ANAL/EXOTIC/CDFR/4348, June 15th, 1999, and G. Apollinari, I. Fiori, P. Giromini, F. Happacher, S. Miscetti, F. Ptohos, *Observation of semileptonic decays of heavy quarks produced in* $p\bar{p}$ *interactions and the existence of one scalar quark*, CDF/ANAL/TOP/CDFR/4795, Feb. 6th, 2000.

The result can be easily guessed: the fit will have no trouble in picking a small upward or downward fluctuation in the number of events, using the additional parameter of the gaussian (the mass) to accommodate it. It is very unlikely that no gaussian bump will be fitted, since there has to be a fluctuation off the generating background distribution somewhere in the spectrum! In fact, from first principles one could guess what the probability distribution for the fitted number of signal events looks like. It has to be a function approximately symmetrical around zero, since the number of events in the distribution is large and the Poisson fluctuations are perfectly gaussian-like. It also has to have asymptotes at zero for $N \to \pm\infty$, reflecting the asymptotic behavior of Poisson fluctuations. But it also must vanish at zero, because of the need of the fit to make good use of the gaussian degrees of freedom. Having no additional features, this function must then possess two maxima, at $N = \pm N_{max}$. The absolute value of N at which these maxima occur can be guessed by assuming that the background is flat: in that case, if the histogram is made of $53,242$ events and if the boundaries are at 6300 and 9000 MeV, there must be on average $N_{ave} = 53,242 \times 76/2700 \sim 1500$ events in any 76 MeV window. These will typically fluctuate up or down by $\sqrt{N_{ave}} \sim 40$, but the fit will actually return a signal larger than that, being able to adjust the mass to pick the most profitable gaussian bump. So we expect to see probably more than 50 fitted signal events on average in a fit. The extent to which the fit allows the mass to float will determine by how much this number will grow: a fit where the mass floats only until a local maximum is found for the total likelihood will find more or less $50 - 60$ events on average, while a fit which forces a scan of the whole spectrum to find the best bump will typically return a larger value.

A distribution like the one just described is shown in Fig. 2. The depletion at zero fitted events is evident, and we like to stress here that it just reflects our allowance for a floating mass in the fit: a fit with mass fixed to a particular value would yield a perfectly gaussian probability distribution for the number of fitted events, as is easy to understand since gaussian is the behavior of Poisson fluctuations in the number of events around their mean value (see Fig. 5). We will discuss in more detail some implications of this observation in Sec. 4.

Now let's move on to discuss another feature of likelihood fits to a spectrum like the one shown in Fig. 1. The large number of background events under a $\sim$ 200-event peak will make it very hard for a real signal to be preferred over some other fluctuation happening somewhere else in the spectrum; this happens if an input value for the mass which is far from the "real" peak is given to the fit, because the fit will quickly content itself with a local maximum in the likelihood function, and will fail to explore all the possibilities in the spectrum. This behavior complicates slightly our task of finding a proper definition of probability for the peak observed in the dimuon data, because we wish to explore the situation where the experimenter was not able to predict a priori and with precision the mass value at which he or she would find a peak in the given distribution. Remember, it is exactly this situation that makes us doubt the simple estimation given by the "pseudo-significance" of the fitted signal in CDF-4860 –computed as $N_{fit}/\sigma_{N_{fit}}$– in the first place.

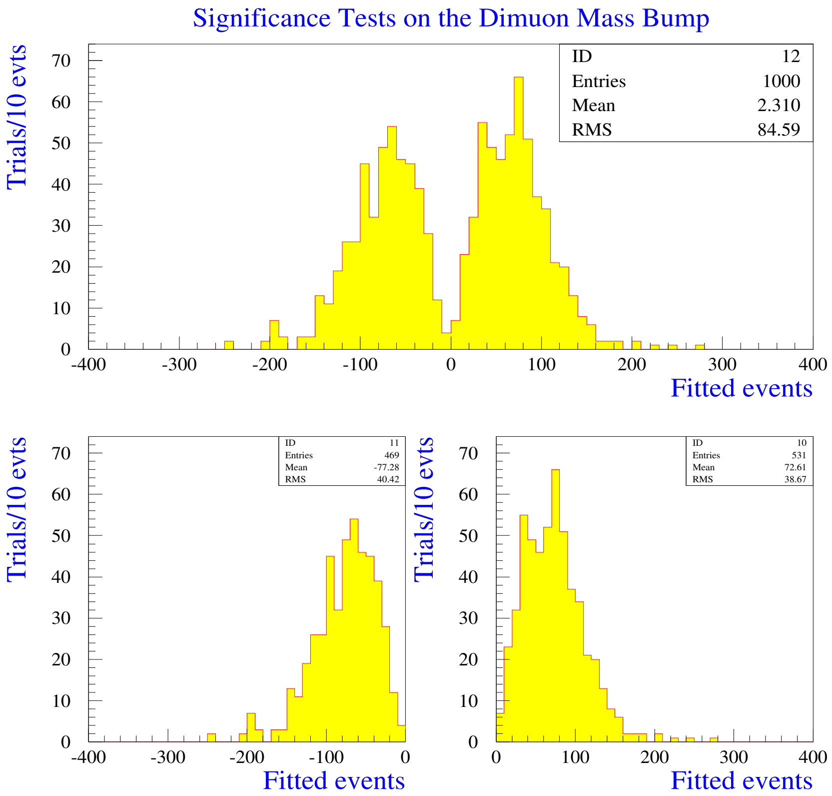

Figure 2: *The number of signal events returned by* 1000 *likelihood fits to spectra of* $53,242$ *events between* 6.3 *and* 9.0 GeV *generated according to a* $exp(-0.5M)$ *distribution, when the width of the gaussian modeling the searched signal is fixed to* 38 MeV*, the mass is allowed to float, its input value is chosen randomly, and no signal is present. The bottom plots show separately the negative and positive tails, in order to highlight their average value and total statistics.*

In the following we will call pseudo-significance the signal significance computed by just dividing the fitted number of events by its relative uncertainty.

In order to obtain a proper estimation of the probability that a signal of the observed size would be found if no signal were present, we must "simulate" not just the experimental spectrum, but the experimental hand as well. We have, that is, to find a procedure by which we choose the input mass in the likelihood fit to a generated mass spectrum as the experimenter did with the real spectrum. Here we are exposed to easy criticism, because of the intrinsic arbitrariness of whatever procedure we may choose, but we will boldly propose one nonetheless. If we argue that the way an input value is chosen for the mass in the real world is by looking at the spectrum, binned in –say– 50 MeV bins, and choosing the mass where two adjacent bins show an upward deviation from an otherwise smooth spectrum,

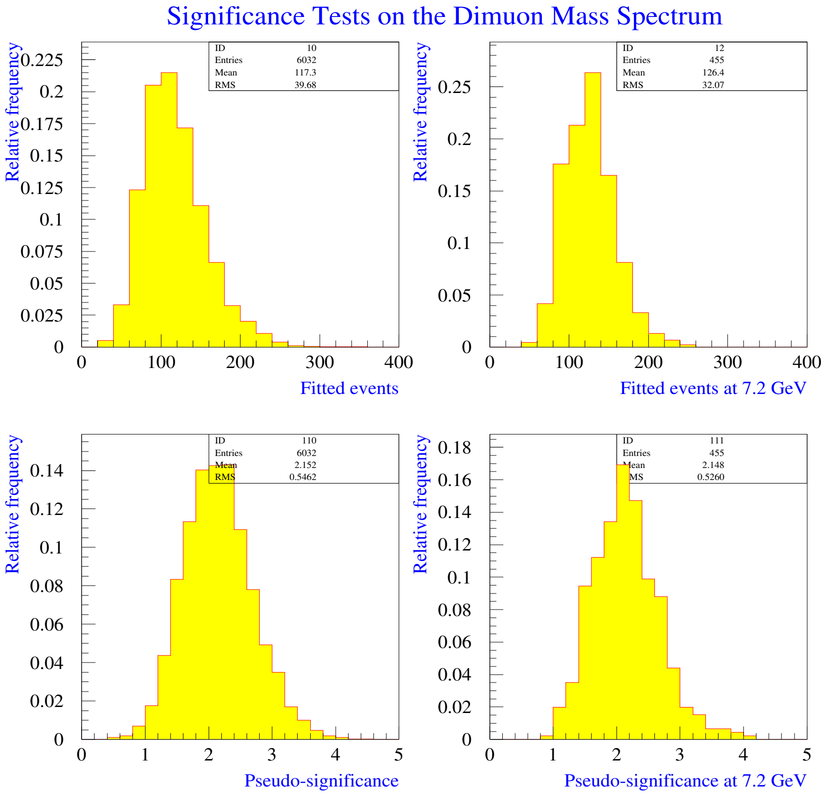

Figure 3: *Upper left: Distribution of the number of fitted events when no signal is present, the mass is left free to float in the fit, its input value is chosen at the point where the spectrum to be fitted shows the most significant upward fluctuation in two adjacent* $50\ MeV$ *bins, and the width is fixed at* $38\ MeV$*. Upper right: Same for pseudoexperiments whose returned mass is within* $0.1\ GeV$ *from* $7.2\ GeV$*. Lower left: The "pseudo-significance" (see text) of the fitted number of events in the upper left plot. Lower right: same for pseudoexperiments whose returned mass is within* $0.1\ GeV$ *from* $7.2\ GeV$*.*

we won't be too far away from the truth. A similar ansatz could have been to choose the place where a single bin is sticking up the most, or where the significance of a single upward point is highest. These should make no real difference for our purpose. We can simulate the whole procedure by simply inserting a further step in our toy Monte Carlo: after generating our $53,242$ events of background, we bin them in $50-MeV$ intervals, and subtract the generating distribution from the resulting histogram. A plot of lone-standers wiggling around zero is thus obtained, from where the two adjacent bins can be chosen whose total pseudo-significance (computed as $S = \sqrt{N_i + N_{i+1} - f(i) - f(i+1)}$, where $f(i)$ is the value of the properly normalized generating function at the middle of bin i) is the highest. The input mass will then be chosen at the center of the pair of bins.

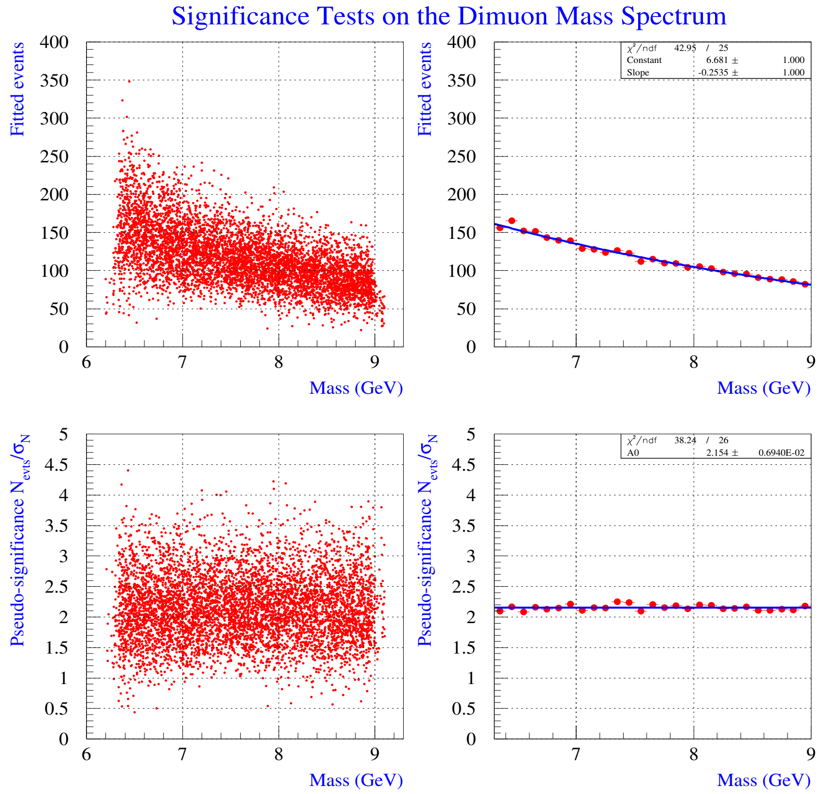

Figure 4: *Upper left: correlation between the fitted mass and the number of fitted events, for pseudoexperiments containing no signal, fitted with fixed width and input mass set at the value showing the most significant upward fluctuation. Upper right: Average number of fitted events as a function of the returned mass; an exponential fit is overimposed. Lower left: correlation between the fitted mass and the pseudo-significance of the number of fitted events (N_{fit}/σ_N), for the same pseudoexperiments described above. Lower right: average pseudo-significance as a function of the returned mass.*

If we now generate a large number of pseudoexperiments, and fit them as the sum of an exponential (whose slope is allowed to float in the fit, as would happen in the real life) and a gaussian, with 38 MeV width and mass free to roam, we obtain a very useful result: a distribution of the number of signal events we would fit in the dimuon spectrum with a unbinned likelihood technique, if our data contained no signal capable of making a difference. The distribution is shown in Fig. 3. We can use the upper left plot in that figure to get a rough idea of what is the typical number of events under a gaussian signal one may get by fitting a spectrum like the one we have from run I dimuon pairs. The lower left plot instead can give a more precise estimation of the real probability that what we saw comes from a background fluctuation: we can, that is, fit the real data as we do with the pseudoexperiments, obtain the pseudo-significance $s = N_{fit}/\sigma N$, and compute $P(s)$ as the integral of this distribution from s to infinity, once it has been normalized. The plots on the right in Fig. 3 show, for comparison, what is the typical distribution of number of fitted events *when the returned mass is close to* $7.2\ GeV$, and what is the relative pseudo-significance. The distribution of N_{fit} is different from the one on the left, due to the dependence of N_{fit} on the fitted mass (Fig. 4), but the significance must be the same, as the lower right plot confirms.

In this first version of this note we will refrain from quoting a probability that the dimuon background produced a peak equal or larger than what is found in the data, deferring that to after a consensus has been reached on the details of the correct procedure to use for the estimation. Such details include defining the mass range where a quarkonium signal can be called a confirmation of the inclusive lepton excess, defining the ansatz for the choice of the input signal mass in a pseudoexperiment, and choosing other details of the fit (kind of statistical estimator –likelihood or chisquared–, bin size –if it applies–, boundaries, background shape, and so forth).

3 Conclusions

The probability that the dimuon bump described in CDF-4860 is produced by a mere fluctuation of the background cannot be precisely computed by a simple-minded look at the ratio between the number of fitted signal events and its error. A precise estimate can be obtained instead with the method described in the present note. The estimation will be carried out after a consensus has been reached among the godparents and the authors on the correct way of taking into account some details of the computation and of the analyses described in CDF-4348, CDF-4795 and CDF-4860. This "blind analysis" approach is, we believe, the only way of treating with an otherwise too controversial issue.

The studies discussed in the present note purposedly avoid addressing some of the "side-issues" concerning the fits to the dimuon mass bump. In particular one issue has remained slightly controversial to the present day, namely the observed width of the signal at $7.2\ GeV$. Unbinned likelihood fits that leave the width of the signal free to vary have been shown by G. Velev to reconstruct a smaller signal than that quoted by the authors of CDF-4860, with a width that is returned to be $14 \pm 5\ MeV$. Is this number inconsistent with the expected $38\ MeV$ due to the momentum resolution for the muon tracks ? It is our opinion that this question cannot be answered by just a look at the returned value for the error on the width. Again, we favor a pseudoexperiment approach.

4 Appendix: The Greedy Bump Bias

4.1 Introduction

To check that our likelihood technique was working correctly we generated a pseudoexperiment containing 100 times the statistics of the real data: $5,304,200$ background and $20,000$ signal events in the $6.3-9.0$ GeV range, with the signal sitting at 7.2 GeV with a width of 38 MeV. By fixing the signal width to 38 MeV in the fit we obtained $N_{fit} = 20,188 \pm 592$ and $M_{fit} = 7.20271 \pm 0.00151$ GeV, while the slope of the exponential used to model the background was generated at $s = 0.5/GeV$ and was fit to $s_{fit} = 0.49959 \pm 0.00058/GeV$. When the same distribution was fit with a floating width the results became $N_{fit} = 20,243 \pm 724$, $M_{fit} = 7.20269 \pm 0.00153$ GeV, $w = 0.0382 \pm 0.0015$ GeV, and $s = 0.49959 \pm 0.00058/GeV$. These numbers reassure us about the reliability of our fitting procedure.

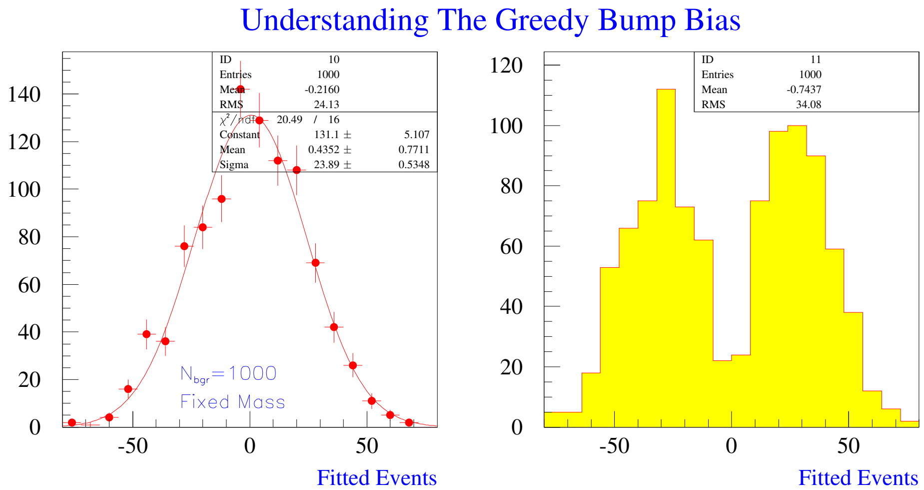

Figure 5: *The number of signal events returned by* 1000 *likelihood fits to spectra of* 1000 *events generated in the interval* $-5 \div 5$ *according to a flat distribution; the signal fit is a gaussian of unit width. Left: the probability distribution for the returned number of signal events is gaussian when the mass of the signal is fixed at zero. Right: allowing the mass to float in the fit will create a double-winged distribution.*

4.2 The Greedy Bump Bias Unveiled

The other obvious, common, and –normally– convincing check one typically performs to verify that a program performing a fit is working properly consists in obtaining a distribution for the so-called "pulls" in the degrees of freedom of the fit. In presence of well behaved

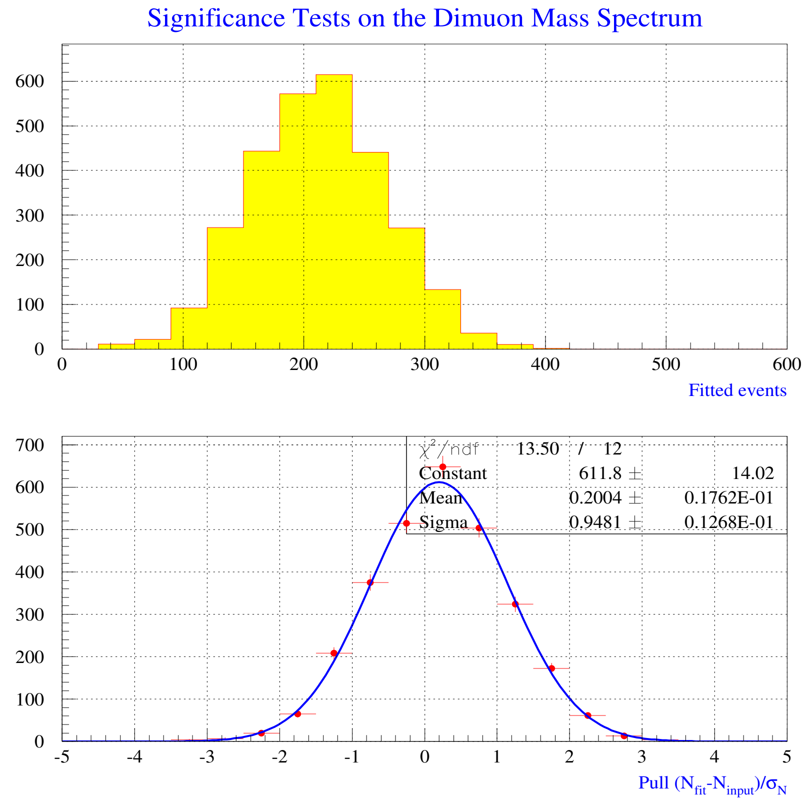

Figure 6: *Top: Distribution of the number of fitted signal events for pseudoexperiments where* 200 *signal events are generated with a mass of* $7.2\ GeV$ *and a width of* $38\ MeV$*, along with* $53,042$ *background events generated according to a* $exp(-0.5x)$ *shape. In the fit the width is fixed at* $38\ MeV$*, while the mass is left free to float and its input value is set at* $7.2\ GeV$*. Bottom: Distribution of the pull in the number of fitted events,* $(N_{fit} - 200)/\sigma_N$*. A bias towards* $+0.2\sigma$ *is evident.*

and gaussian errors these pulls, defined as the number of standard deviations from the true value of a parameter returned by the fit, should distribute normally (i.e. as a gaussian of zero mean and unit r.m.s.) if the fit is working properly and there are no biases.

In the presence of a large background and a small signal *whose mass is allowed to float in the fit*, the number of fitted events will NOT, on the contrary, distribute normally around the input value. This is a general result and is worth describing in detail.

Let's consider the distribution of the number of fitted events in the absence of a signal, when the mass is left free to vary within the histogram boundaries. We showed that distribution in Figs. 2 and 5: its features are quite general, and vary only slightly as a function of the kind of background shape and normalization. For instance, if the histogram to be fitted contains only a few events, the large negative tail of the distribution of N_{fit} will be suppressed, since downward fluctuations become unlikely if not impossible; the fraction of times that, say, a positive signal is returned will also depend slightly on the shape of the distribution. In light of these considerations, the following discussion retains some general value, as will be made clear shortly.

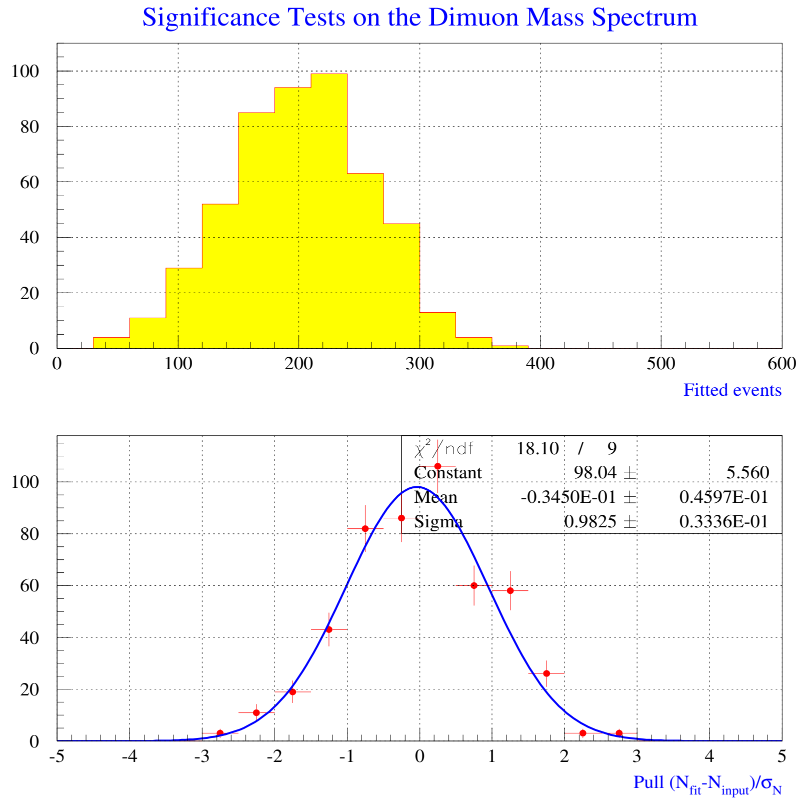

Figure 7: *Top: Distribution of the number of fitted signal events for pseudoexperiments where 200 signal events are generated with a mass of 7.2 GeV and a width of 38 MeV, along with 53,042 background events generated according to a* $exp(-0.5x)$ *shape. The width is fixed at 38 MeV, and the mass is fixed at 7.2 GeV. Bottom: Distribution of the pull in the number of fitted events,* $(N_{fit} - 200)/\sigma_N$. *There is no sign of bias, as expected.*

We have already noted that for zero signal events the probability distribution of the number of fitted events will have the shape of Fig. 2: this is a highly non-normal distribution, and so will be the distribution of the pull N_{fit}/σ_N, of course, regardless of the number of events that populate the fitted histogram. But zero signal events is not a general case, so let's consider a one-event signal. No kidding: throwing one signal event will make practically no difference, unless the background itself is really small. The pull distribution will retain its double-winged shape; only, it will shift slightly to the right and the positive wing will gain slightly in normalization with respect to the negative wing. Let's throw in more signal events: intuitively, the distribution will gradually shift to the right, deflating its negative tail, and starting to resemble more and more a gaussian in the positive wing. How long will this go on ? Well, it depends entirely –to first order, at least– on the significance ratio $S/\sqrt{S+B}$ between the number of events in the gaussian peak and the amount of background below it. Why ? Because until Poisson fluctuations on the number of background events under the peak stop being comparable in size with the signal itself, their contribution to the number of fitted events will be felt strongly, just as it can be intuitively understood considering the transition from a Poisson-like shape to a gaussian shape. Pseudoexperiments show that this effect is still quite strong for 200 signal events over a background of about 2500 sitting underneath it, as is the case for a 38 MeV resonance in our familiar case of the dimuon spectrum[6]: the pull will still be offset by 0.2 σ ($\sim$ 10 events), which is a fat 20% of the typical fluctuation of the 2500 background events (Fig. 6).

If both background and signal are reduced by a factor of ten (5324 vs 20 in the whole fitted range, $6.3 \div 7.0$ GeV) things become much worse, because the fluctuations in the background are in this case $\sqrt{10}$ times stronger with respect to the size of the signal: in fact, we observe a shift of 0.7σ on the number of fitted events (see Fig. 8). We've dubbed this problem the "Greedy Bump Bias", since the bump really appears to prefer eating out positive fluctuations than withstanding negative ones.

We fear we've not convinced all of our seven readers with these intuitive arguments, even if the few plots we have thrown in tell a lot of words themselves. There is actually another way of understanding why a decent, reasonable likelihood fitting technique is going to prefer seeing more signal events than we feed to it. Let's turn our mind to the original background spectrum with the small gaussian peak on top of it, Fig. 1. Why will the fit prefer a positive fluctuation of the background to a negative one ?

The answer can be put in the following words: *because a positive fluctuation makes the fitted signal more robust*. Since the error on the fitted signal will be dominated by the size of the background, which in our familiar example does not change and is of the order of 2500 events under a 38 MeV gaussian bump, the error on N_{fit} itself is constant and ignores the size of the signal; thereafter, the "pseudo-significance" of the fitted signal, or N_{fit}/σ_N, ends up depending *linearly* on N_{fit} itself. And N_{fit} CAN be made larger by the fitter, which opportunistically moves the fit mass to accommodate a small upward fluctuation in the background, occurring close to the true signal mass. Why, you may ask, what does the fit care about the gaussian bump being more robust ? After all, all the fit knows is the total chisquared, or the total likelihood, of a given set of parameters. Well, this is technically true, but through the chisquared the fit does get to know about how the gaussian bump is doing. In the case where there is no signal, in fact, the fit benefits from the most significant upward or downward fluctuation being picked up by the gaussian degrees of freedom. This behaviour will continue even in the presence of a real signal, as long as the mass is allowed to float. Of course, if the mass is fixed in the fit, the biases disappear completely (see Fig. 7).

[6] 2500 is roughly the number of events in a 100 MeV window around 7.2 GeV, which will therefore contain most of the signal.

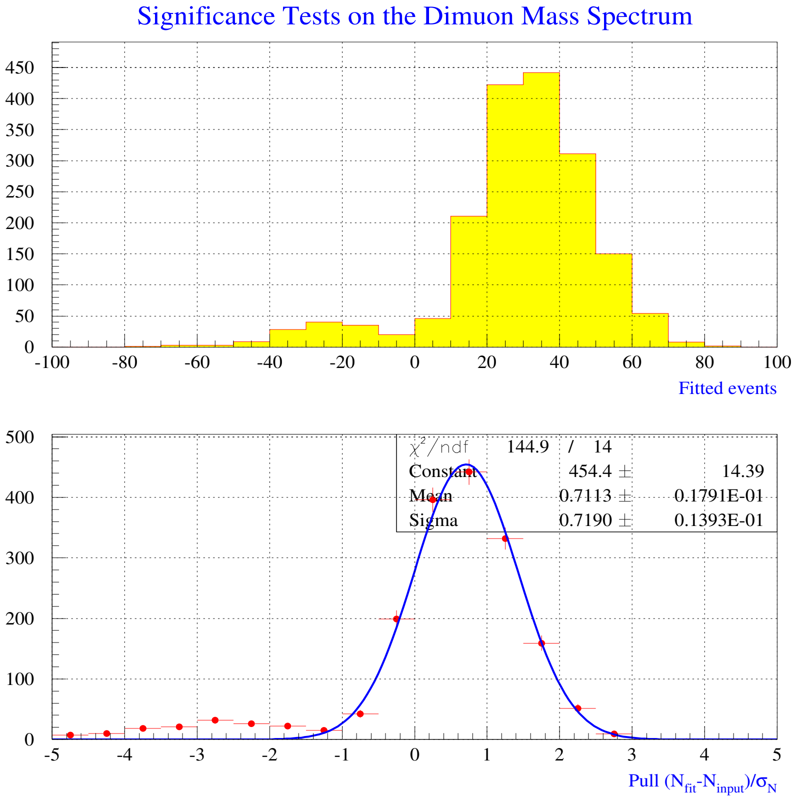

Figure 8: *Top: Distribution of the number of fitted signal events for pseudoexperiments where* 20 *signal events are generated with a mass of* 7.2 GeV *and a width of* 38 MeV*, along with* 5304 *background events generated according to a* $exp(-0.5x)$ *shape. The width is fixed at* 38 MeV*, the mass is left free to float, and its input value is set at* 7.2 GeV*. Bottom: Distribution of the pull in the number of fitted events,* $(N_f it - 20)/\sigma_N$*. The bias is very strong, and a remnant of the negative tail of the distribution is also evident.*

Another way of seeing it is the following: being confronted with the choice to fit a large signal in one point of the spectrum and a smaller signal in another point, the fitter will prefer the larger one, because the larger signal is the one most offset from the background shape, so that the chisquared will get a higher benefit. Again, this freedom of choice is canceled when the mass is fixed.

This is the “Greedy Bump Bias”. Amazingly, it has no cure. A fit with a floating mass will suffer from this bias, regardless, to first order, of the background shape. The bias will depend on the size of background and signal, period.

4.3 Estimating the Greedy Bump Bias

At this point, one would like to estimate the size of this bias once and for all. If we have to live with it, we may just as well determine its value and subtract it back every time we have to fit a small signal with a large background, if we have no excuse to fix the mean value of the signal. Well, things are not so easy, unfortunately. The problem is that, until the bias becomes totally insignificant, there is always a small negative tail in the number of fitted events, the remnant of the double-winged shape we have first met in Fig. 2. In the real life we tend to forget about this negative tail, since when we encounter a negative fitted signal we are likely to deem it a failure of the fit to converge, and proceed to change the background shape, allow the width to float, fix the mass, or –God forbid– rebin the data! Now, how to deal with all that is a problem much bigger than just calling off the negative tail and proceed to estimating the bias in a very simple way, with the help of our toy Monte Carlo. That's what we'll try to do in the following.

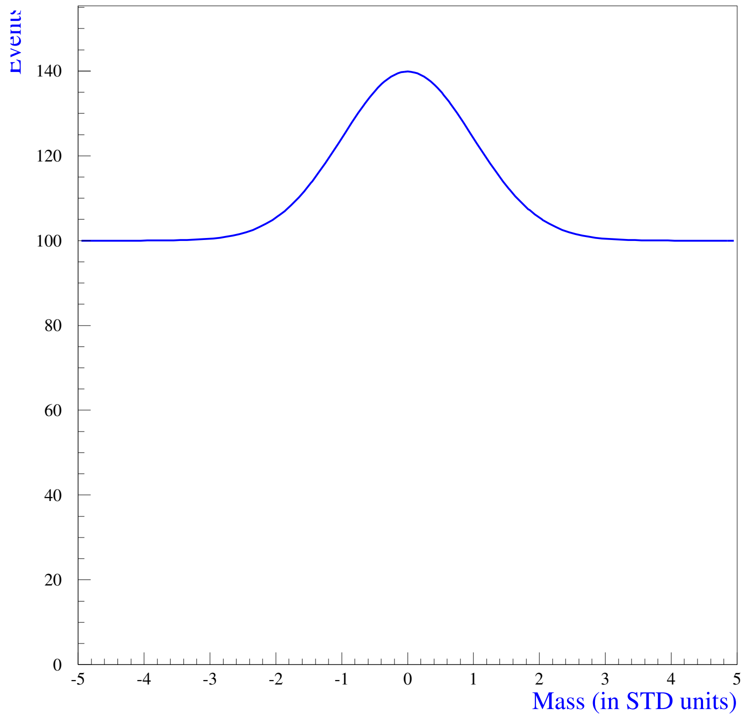

Figure 9: *A standardized case to study the greedy bump bias: a gaussian signal of unit width and zero mass, consisting of* 100 *events, sits on top of a flat background of* 1000 *events.*

We take a very simple situation into consideration: a gaussian signal on top of a flat background, as in Fig. 9. It can be argued that the results will depend only very slightly on the actual shape of the background, provided that the latter does not vary significantly in the range where the gaussian signal sits.

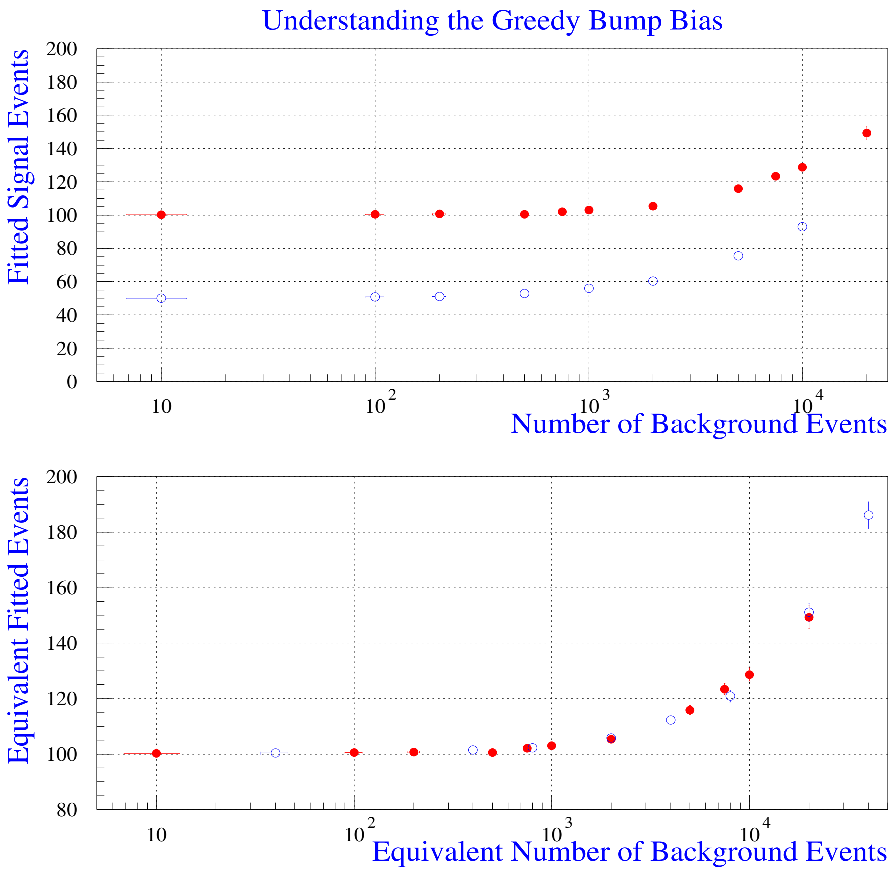

Figure 10: *Top: the average number of fitted signal events is a monotonous function of the amount of background events. The two sets of points have been generated from pseudoexperiments with $N_s = 100$ (full red dots) and $N_s = 50$ (empty blue dots). Pseudoexperiments returning a negative number of signal (a not infrequent instance in the high background points) are used in the computation of the average. Bottom: The two sets are compared by rescaling $N_s' = 2 * N_s$ and $N_b' = 4 * N_b$, in order to verify that the bias scales with the ratio $S/\sqrt{B}$. The above conjecture is borne by the data.*

A signal of $N_s = 100$ events, of mean zero and unit width, is generated along with N_b background events in the range -5 to $+5$ (Fig. 9). The resulting spectrum is fit with our usual unbinned likelihood technique to the same functions used to generate the events, by keeping the width fixed at 1.0 and allowing the mass to float. The input values of all free parameters are set to their respective true value. The fitted number of events is then stored along with its uncertainty, and the whole procedure is repeated 1000 times. Then N_b is varied to investigate another situation. The whole procedure outlined above is then carried out again, starting this time with a signal of only $N_s = 50$ events to investigate the supposed "universality" of the bias.

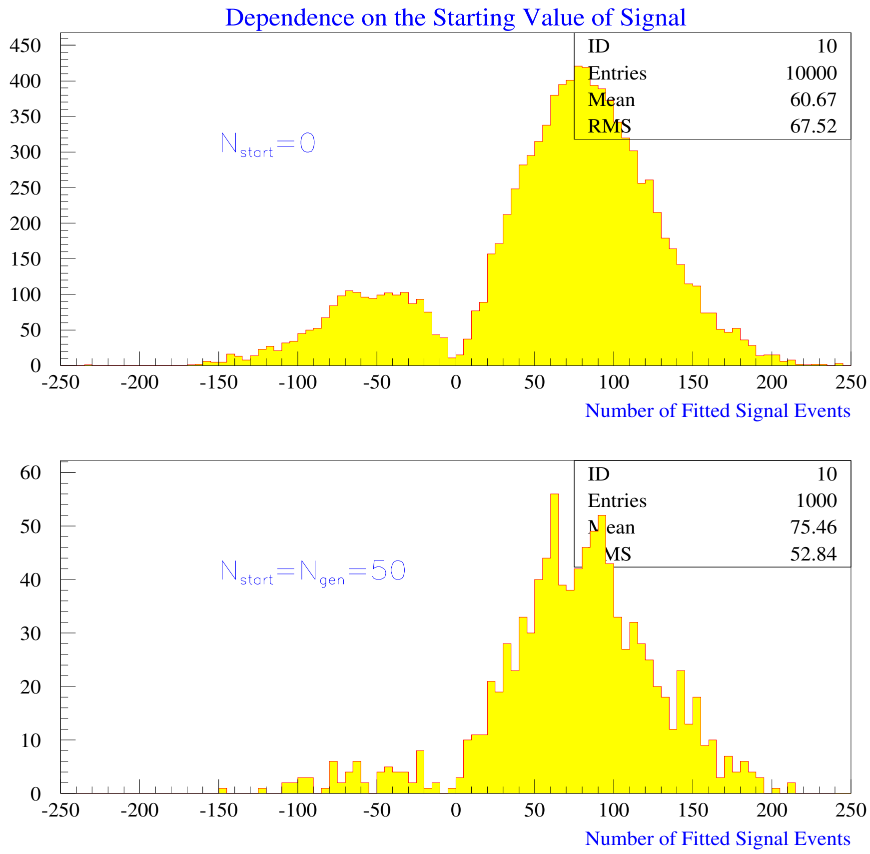

Figure 11: *Top: the average number of fitted signal events for* $N_s = 50$, $N_b = 5000$ *shows a* 20.1% *bias if the fitter is initialized with zero signal events, while it shows a bias of* 50% *if the fitter starts with* $N_s = 50$.

We can finally compute the distributions of the number of fitted signal events, and plot their mean as a function of N_b, for the two cases $N_s = 50$ and $N_s = 100$. We believe the mean values of these distributions are more informative than the position of the peak of gaussian fits to either the distributions of the number of fitted events or the distributions of their pulls, since these become quickly non-gaussian when the amount of background becomes large. The results are shown in Fig. 10: one clearly sees that the bias becomes rapidly significant as $S/\sqrt{N}$ increases.

The monotonic behavior shown in Fig. 10 is unsuspicious at first sight, but counterintuitive on a second thought: if one starts wondering what happens when the number of background events becomes enormous, things get confusing. In fact, when the signal is totally insignificant, why should it matter anymore ? One would then think that again a symmetric double-winged distribution can be obtained for the number of fitted events, whose mean cannot show a bias anymore. In fact, there seems to be another parameter governing the behavior of the bias. Minuit requires the user to provide a starting value for every parameter in the fits: the number of signal events is one of these. In the fits producing Fig. 10 this number has been set to the actual number of generated signal events in the pseudoexperiments. What happens instead if this number is set to zero, allowing the fit to decide by itself whether a positive signal is there or not ? The answer is that the distribution of the number of fitted events changes its relative normalization between its positive and negative wing (Fig. 11).

That is, the shape of each wing does not seem to vary, but the fraction of times that a negative signal is found increases: this amounts to a smaller bias, that does not appear to blow up as the signal significance decreases: this more “reasonable” behavior is shown in Fig. 12, where the bias is compared for the two cases. The bias does indeed go away for very insignificant signals, provided that the input value of the signal sought is set to zero. This is again a remarkable observation: the average of the fitted number of events, in fact, does not seem to converge to zero for $S/\sqrt{B} \to \infty$, rather it seems to converge to S, even if S is really insignificant. We think this is nothing by yet another manifestation of the central limit theorem (see Fig. 13).

More investigation is needed in order to understand better the behavior of the greedy bump bias.

We believe that the greedy bump bias leaves unaffected most of CDF's b physics results, that typically deal with signals of high pseudo-significance; nevertheless we suggest all analyses measuring the cross section of small signals to perform their own pseudoexperiments, in order to evaluate if, and by how much, they are overestimating their signals in their particular case.

Acknowledgements

Many thanks to A. Ribon who, besides being the original author of the code used for the fits, has been very helpful with many brainstorming discussions.

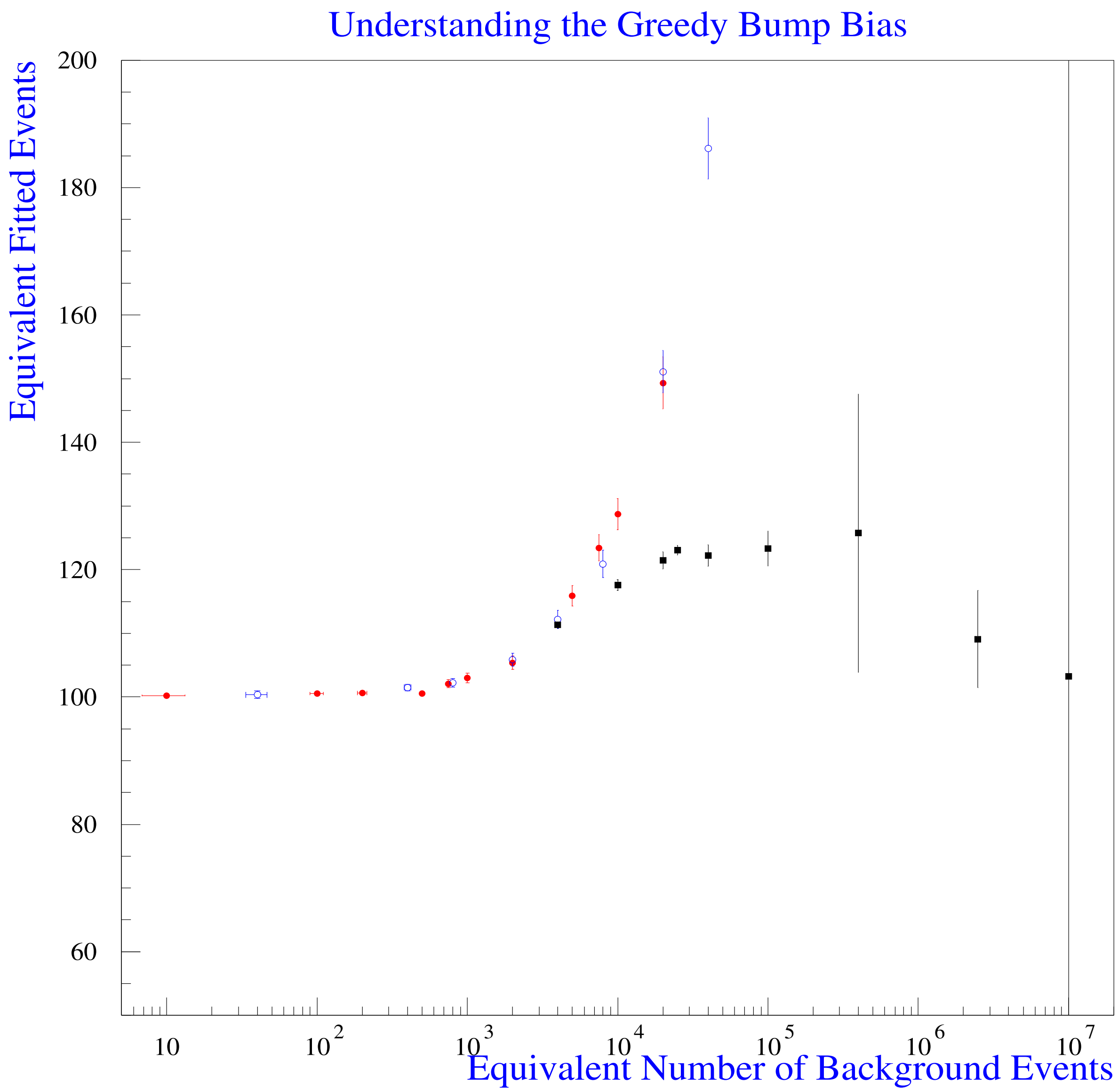

Figure 12: *The black squares show the behavior of the bias (computed with different combinations of the number of generated signal –1, 2, 5, 10, 20 and 50– and background –1000, 5000, and 10000– events) rescaled to the equivalent number of fitted events for a generated signal of 100, using the hypothesized constancy with $S\sqrt{B}$; Minuit is initialized with a signal of zero events. The red and blue points, diverging when B becomes large, have instead been obtained by setting $N_{start} = N_{gen}$ for the signal.*

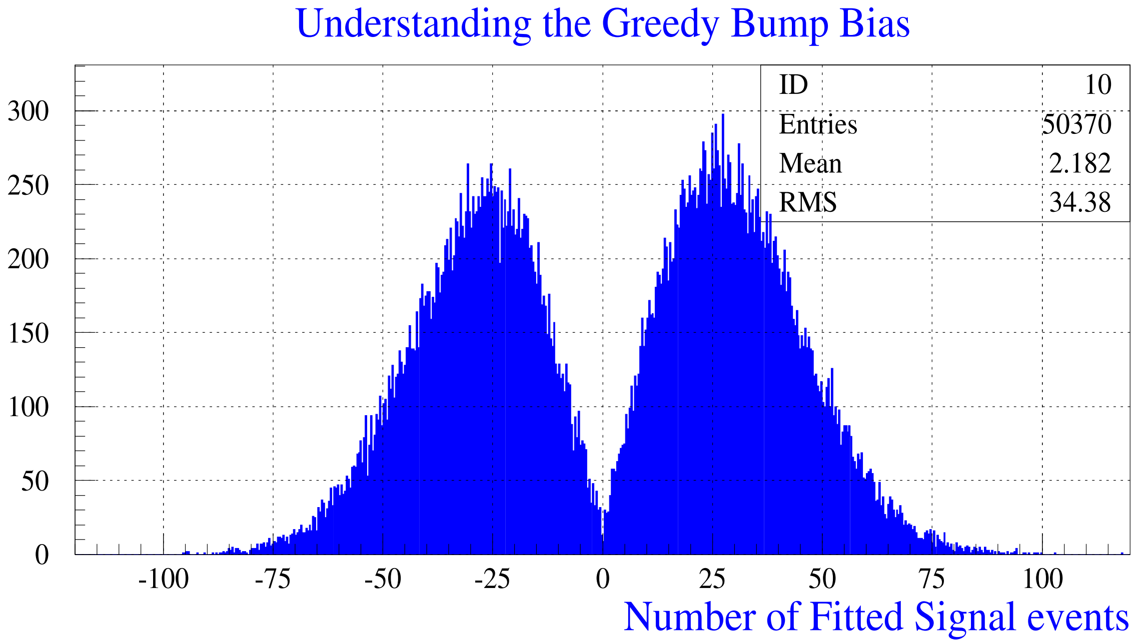

Figure 13: *Distribution of the number of fitted signal events when* 2 *signal events have been generated on top of a background of* 1000*, and the fitter is initialized with a starting value of zero signal events. Although the signal is totally negligible, the average of this wide distribution converges remarkably to the number of signal events in the samples.*