Optimal Bin Size vs. Nexp and Nfit Cut for Milagrito

by **J. F. McCullough**

February 15, 2000

Abstract

I have completed a detailed Monte Carlo study of the optimal bin size to use when doing a binned source search. I started by investigating the optimal bin size as a function of Nexp assuming a detector with Gaussian angular resolution. I then looked at the optimal bin size as a function of Nexp and nfit cut using our 'real' resolution for Milagrito. The optimal bin size for Milagrito when the source position is known is given by :

bin size (in degrees) = 2.96 - 0.53 $\times$ log10(Nexp) + 0.05 $\times$ log10^2(Nexp)

where Nexp is the number of expected background events in a 2^o square bin. If the source position is unknown, the optimal bin size (assuming oversampling by 1/2 a bin) is given by :

bin size (in degrees) = 3.36 - 0.67 $\times$ log10(Nexp) + 0.08 $\times$ log10^2(Nexp)

1 Procedure for a Detector with Gaussian Resolution

I began this study by first investigating the optimal bin size to use (as a function of the number of expected background events - Nexp) for a detector with Gaussian angular resolution. This was done mainly to serve as a test of my code. I wanted to reproduce the results from Alexandres et. al., 1992. which found that the optimal bin size for large Nexp was a round bin with radius 1.58σ (where σ is the angular resolution) or a square bin with sides $\pm$1.4σ in declination(δ) and $\pm 1.4\sigma/\cos(\delta)$) in right ascension. I looked at the optimal bin size using both round bins and square bins. The procedure I used is as follows:

1) A 20$\times$20^o square box was created in which to throw background and signal events (see Figure 1.).

2) I choose 18 different Nexp from 0.10 to 1000.0. When using square bins, Nexp was defined as the number of expected background events in a square bin with sides of length 2σ. When using round bins, Nexp is the number of expected background events in a round bin of radius σ.

3) The appropriate number of background events (NBACK) was then thrown in the 20$\times$20^o square box. When using square bins, NBACK = Nexp $\times$ (20^2 / (2σ)^2). For round

bins, NBACK = Nexp $\times$ (20^2 / $\pi\sigma^2$). This is Nexp normalized to the big box. This large of a throw area should account for fluctuations in the background for a specific Nexp.

ex. For an Nexp of 5.0 and an angular resolution of σ=1.0°, I threw a total of 500 events (randomly from a uniform distribution) in the 20° square box when investigating optimal square bins (See Figure 1.). For round bins, I threw a total of 637 events.

4) The number of signal events to throw, NSIG, was determined by figuring out how many observed events (Nobs) would be required to give a signal significance of 10^{-4} in the biggest bin size investigated. For square bins, the biggest bin size I looked at was 6.0×6.0°. For round bins, the biggest bin size I looked at had a radius of 3.5°. NSIG was then given by the following:

NSIG = Nobs - Nexp (where Nexp is normalized to the biggest bin)

ex. For an Nexp of 5.0 and σ=1.0° using square bins, Nexp normalized to a 6×6° square bin is 45 events. An Nobs of 73 events (when I expect 45) gives a Poisson probability of 7.68×10^{5}. I therefore threw 28 signal events in this case (see Figure 1.).

5) NBACK events were thrown randomly from a flat distribution.

6) NSIG events were then thrown from a 2D Gaussian (σ equal to the angular resolution) centered in the middle of the 20×20° square box.

7) The background and signal events were binned in 0.05° square bins.

8) The number of background and signal events were then counted in the different bin sizes from 1.0×1.0° to 6.0×6.0° square bins in 0.1° steps. (For round bins I went from 1.0° to 3.5° radius round bins in 0.10° steps). All bins were centered in the middle of the big box.

9) For each bin size, the significance of the source detection was then calculated from Poisson statistics: significance = PoissProb(Nsignal,Nback)

Nsignal is the number of signal + background events in each bin

Nback is Nexp normalized for the bin size

10) Steps 5 - 9 were repeated 10,000 times for each Nexp.

2 Results for Detector with Gaussian Resolution

Figure 2. shows the results of 10,000 trials using square bins, an Nexp of 0.50, and a signal significance of 10^{-4}. For each bin size, the maximum probability and the average probability are plotted. The maximum probability is the probability bin which had the most entries and the average probability is the average of log10 of the probability. As you can see, the average probability gives a nice smooth curve whereas the maximum probability is very jagged because of the quantization of the data (see Figure 3). For this reason, I chose the average probability as the relevant quantity to use for determining the optimal bin size. The bin size with lowest average log10(probability) was chosen as the optimal bin size. Figure

4 shows the results of the average probability vs. bin size for 5 different Nexp (using an angular resolution of 1.0^o and a signal significance of 10^{-4}). As would be expected, as Nexp decreases the optimal bin size increases.

To be sure that the optimal bin size (as a function of the angular resolution) did not depend upon the angular resolution, I looked at three different resolutions: 0.5^o, 0.75^o, and 1.0^o. The results are shown in Figure 5. To within statistical uncertainty, the optimal bin size (as a function of the angular resolution) is independent of the resolution. (ie. optimal bin size in degrees / angular resolution is approximately the same for all three resolutions) The curves are not smooth because I used 0.10^o steps and did not fit the data.

I also wanted to see if the optimal bin size depended upon the strength of the signal. Figures 6 and 7 show the optimal bin size vs. Nexp for three different signal strengths using both round bins and square bins. The optimal bin size does seem to have a slight dependence upon the signal strength. The stronger the signal, the bigger you can make the bin, although this is not a very strong effect. For large Nexp, the optimal bin size for round bins does level off at a radius of $\sim 1.60\sigma$ as expected for Gaussian resolution. For square bins, it levels off at $\sim \pm 1.4\sigma$. These are in agreement with Alexandreas et. al.

All of the previous plots assumed that the source position was known. To investigate the affect of not knowing the source location (such as in a blind GRB search), I repeated the above steps (1-10) for a unknown source position. This was done by offsetting the source by $\pm 1/4$ the optimal bin size from the center of the big 20^o box for each Nexp. For example, for an Nexp of 0.5, a signal strength of 10^{-4}, and square bins, the optimal bin size was found to be 3.8^o. To simulate an unknown source position, the source was then offset by $\pm 0.95^o$ from the center of the big 20^o box. This offset was chosen because any offset beyond this would be centered in an adjacent bin assuming an oversampling scheme where bins are offset by 1/2 a bin (see Figure 8). The results for an unknown source position are shown in Figure 9. The optimal bin size is approximately 10-15% larger when the source position is unknown (note that this results depends upon the oversampling scheme used).

3 Procedure using Milagrito's Angular Resolution

When simulating Milagrito's angular resolution, I looked at the optimal bin size and nfit cut vs. Nexp. I looked at nfit cuts from 0 to 60 in steps of 5. I used square bins and a signal significance of 10^{-4}. This signal significance was chosen because it is better to optimize on a weak signal and get a slightly smaller significance on a strong signal than optimize on a strong signal and possibly miss a weaker signal. Square bins were chosen because that is what is used for the untriggered burst search. One could easily extend the results to round bins by converting to the appropriate area round bin. Nexp was defined as the number of expected background events in a square bin of length 2.0^o. To characterize Milagrito's angular resolution, I decided to use deleo/2 from our data (I used run 139 subrun 58). I broke deleo/2 up into 23 different nfit bands: 0-10, 10-20, ... (see Figure 11). There are certainly problems with using deleo/2 from data to characterize our resolution. Deleo isn't sensitive to systematics such as curvature and using data gives our resolution for proton showers and not gamma showers. However, there were not enough Monte Carlos showers at the time to use delangle or some other quantity derived from Monte Carlo. When this same study is done for Milagro (for which we have sufficient numbers of Monte Carlos), more care will be taken in choosing the 'right quantity' to characterize our angular resolution.

I used the same steps as outlined earlier for a detector with Gaussian resolution except for how the background and signal events were thrown (steps 5 and 6). When throwing the background events, I first determine the percentage of events that survive each nfit cut (see Figure 10). I then throw that number of background events for each nfit cut. For example, only 79 percent of events survive an nfit cut of 30. Therefore when looking at an nfit cut of 30, I only threw 79 percent of the number of background events as I would have normally thrown. When throwing signal events, I first threw randomly from our nfit distribution (see Figure 10). If the random nfit was less than the nfit cut, I threw the event out. If the random nfit was greater than the nfit cut, I use the appropriate deleo/2 distribution with a random phi centered in the middle of the big $20 \times 20^{\circ}$ box. For example, for an nfit cut of 30, if the random nfit chosen was 23, the signal event is thrown out. If the random nfit was 42, r was thrown randomly from the deleo/2 distribution for nfit between 40-50 and phi was thrown from a flat distribution.

4 Results using Milagrito's Angular Resolution

For each Nexp, I figured out the optimal bin size for each nfit cut. As before, the bin size which gave the lowest average log10(probability) was chosen as the optimal bin size for that nfit cut. Figure 12 shows the results of several different nfit cuts for an Nexp of 1.0. As can be seen, there is a fairly wide range of nfit cuts and bin sizes which give approximately the same signal significance. The results for Nexp = 1.0 are shown below:

Nfit cut	Optimal bin size	Probability
0	3.000000	-4.037809
5	2.950000	-4.055160
10	2.900000	-4.053312
15	2.900000	-4.064173
20	2.900000	-4.121840
25	2.950000	-4.125255
30	2.900000	-4.170878
35	2.650000	-4.009905
40	2.750000	-3.835009
45	2.550000	-3.599830
50	2.400000	-3.470633
55	2.350000	-3.265371
60	2.200000	-3.127039

Table 1: Results of optimal bin size and nfit cut for Nexp = 1.0. The optimal binsize is 2.9° and the optimal nfit cut is 30.

For each Nexp, the combined nfit cut and bin size which gives the lowest average log10(probability) is chosen as the optimal nfit cut and bin size. For Nexp = 1.0, the optimal nfit cut and bin size was 30 and 2.9^o respectively. The results for all Nexp as shown in Figure 13. The nfit cuts and bin sizes in parentheses for Nexp of 0.5, 0.75, 2.0, and 500.0 were the ones used when calculating the final optimal bin size function. Even though these were not the optimal nfit cut and bin size, they were used to simplify the results. Using these numbers, the data can be broken up into three regions with nfit cuts of 20, 30, and 40, respectively: Nexp $\leq$ 1.0, 1.0 $<$ Nexp $\leq$ 100, and Nexp $>$ 100.0. Since there is a wide range of nfit cuts and bin sizes that give approximately the same results (see Figure 14), these simplifications do not adversely affect the results.

The fitted results for a source of known position is shown in Figure 15. When simulating a source of unknown position, the source center was offset by up to $\pm 1/4$ the optimal bin size for that particular Nexp and nfit cut. For example, for an Nexp of 1.0, the source center was offset by up to $\pm 0.75^o$ for an nfit of 0 and by up to $\pm 0.55^o$ for an nfit of 60. The fitted results for an unknown source position are shown in Figure 16. Figure 17 shows the results for both a known and unknown source position plotted together. As with the Gaussian resolution, the optimal bin size when the source position is unknown is approximately 10-15% larger than if the source position is known. Again, this result is dependent upon the degree of oversampling used (in this case, the results are for oversampling by 1/2 a bin).

5 Other Studies

I wanted to see how our resolution changed with different water depths. For each water depth (1.0m, 1.5m, and 2.0m), I used one subrun of data to get nfit and deleo/2 distributions. I used an Nexp of 1000.0 and a signal strength of 4σ. For each water depth, I looked at nfit cuts from 30 to 60 in steps of 5. Shown in the top plot of Figure 18 is signal significance vs. bin size for the nfit cut which had the best results for each water depth. The bottom plot shows the same thing normalized for event rate. To normalize the different water depths, I used the following event rates: 1.0m (300Hz), 1.5m (390Hz), 2.0m (450Hz). As can be seen, although 1.5m (with an nfit cut of 45) gave the best angular resolution, 2.0m (with an nfit cut of 50) was the optimal water depth once event rate is taken into account.

In order to see how Milagrito's resolution compared to a detector's with Gaussian resolution, I chose an Nexp of 1000.0 and a signal strength of 4σ. I looked at Milagrito's resolution (for 1.0m water depth) with an nfit cut of 45 and Gaussian resolutions of 1.0^o, 1.25^o, and 1.5^o. The results are shown in Figure 19. If you look at where the curves peak (Note: the y-axis is now sigma and not log10(probability)), Milagrito's resolution is comparable to a detector with Gaussian resolution of $\sim 1.20\sigma$.

For the Markarian 501 paper, I looked at the optimal bin size and nfit cut to use for Mrk 501. I used an Nexp of 2500.0, nfit cuts of 30, 40, 50, and 60, and a signal significance of 4σ. To simulate a signal from Mrk 501, I only used data from the vicinity of Mrk 501 to get nfit and deleo/2 distributions. The analysis was done using square bins and then the results were converted to the appropriately sized round bin (this was done because the results were need quickly at the time and square bins was the quickest and easiest analysis to run). The results are shown in Figure 20. The optimal bin size for Mrk 501 is a round bin with radius 0.93^o and nfit cut of 40.

6 Conclusions

I investigated the optimal bin size and nfit cut to used for a binned search in Milagrito. This was done for the source position known (as in a DC search such as Markarian 501) and unknown (as in a blind GRB search). The results are given as a function of Nexp, the number of expected background events in square bin of length 2^o:

known source position:
optimal bin size (in degrees) = 2.96 - 0.53 $\times$ log10(Nexp) + 0.05 $\times$ $\log10^2$(Nexp)

unknown source position (assuming oversampling by 1/2 a bin):
optimal binsize (in degrees) = 3.36 - 0.67 $\times$ log10(Nexp) + 0.08 $\times$ $\log10^2$(Nexp)

The nfit cuts to be used also depend upon Nexp:

Nfit cut = 20 for Nexp $\leq$ 1.0

Nfit cut = 30 for 1.0 $<$ Nexp $\leq$ 100.0

Nfit cut = 40 for Nexp $>$ 100.0

These results have been used in both the all sky untriggerd burst search and in my triggered burst search (results from both will be in an upcoming memo).

7 References

D.E. Alexandreas et. al., Nucl. Instr. and Meth. A328 (1993) 570-577

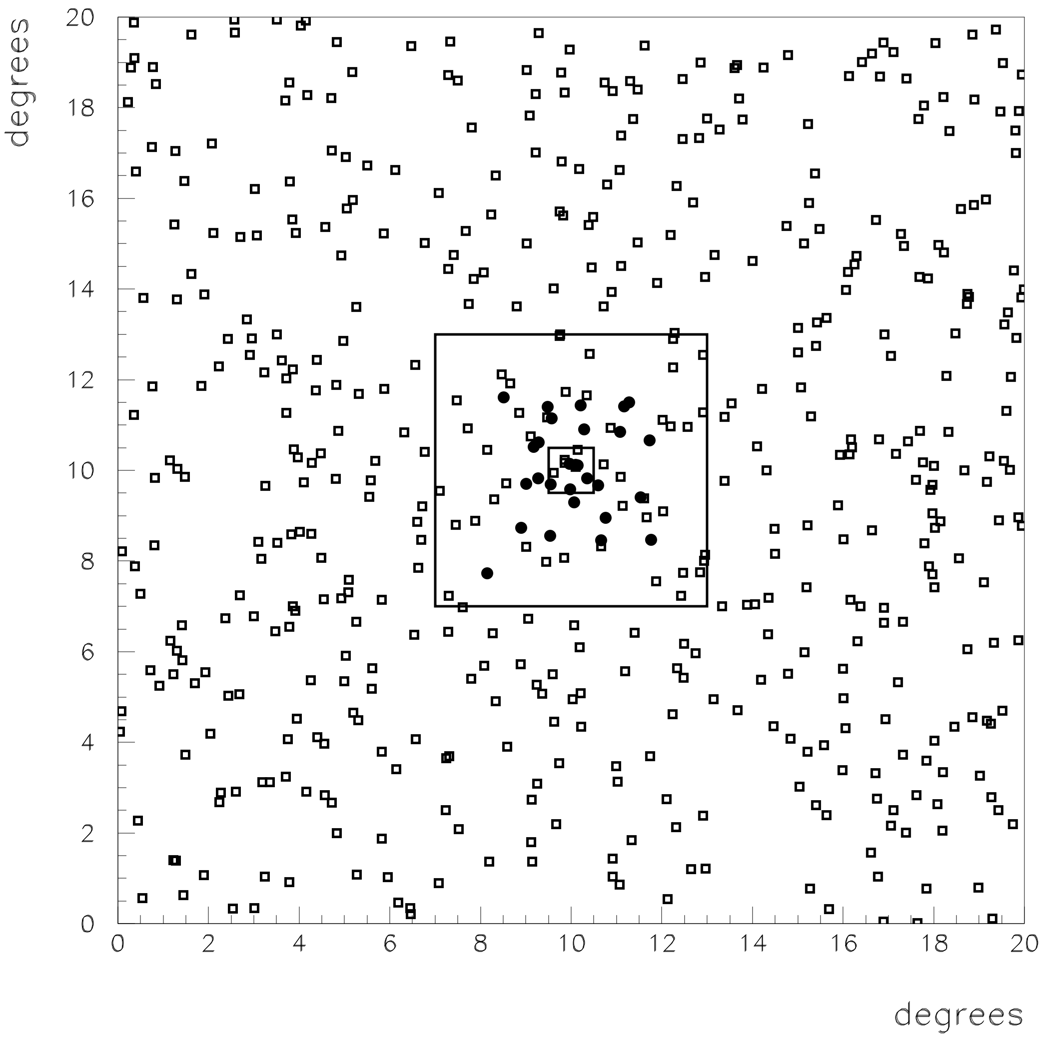

Figure 1: Example of one trial for Nexp = 5.0, σ = 1.0°, and using square bins. 500 background events (squares) and 28 signal events (circles) were thrown. For each trial, the number of signal events (NSIG) and background events (NBACK) were counted in bins from 1.0×1.0° (the small box) to 6.0×6.0° (the big box) in 0.1° steps. For each bin, the signal significance was calculated from Poisson statistics - significance = poissProb(Nsignal,Nexp) where Nsignal = the number of signal + background events and Nexp = Nexp normalized for the bin size.

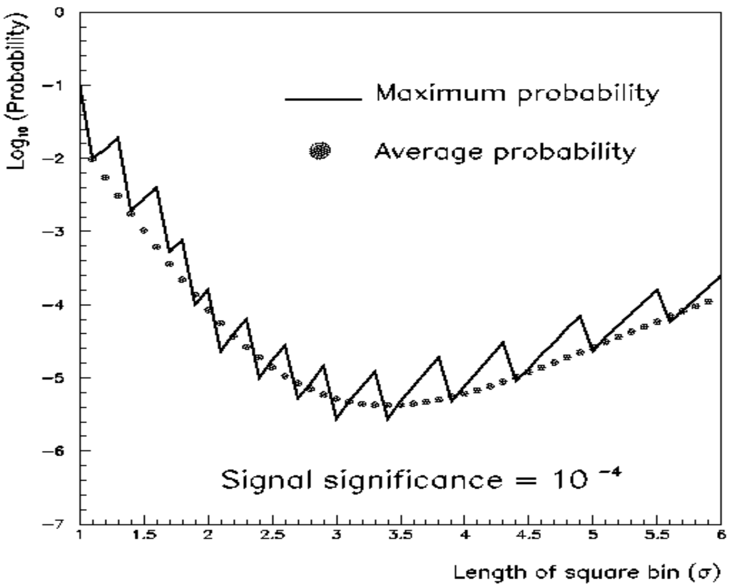

Figure 2: Results from 10,000 trials using square bins, an Nexp of 0.5, and a signal significance of 10^{-4}. The maximum probability for each bin size is the probability bin with the most entries. The average probability is the average of log10(probability).

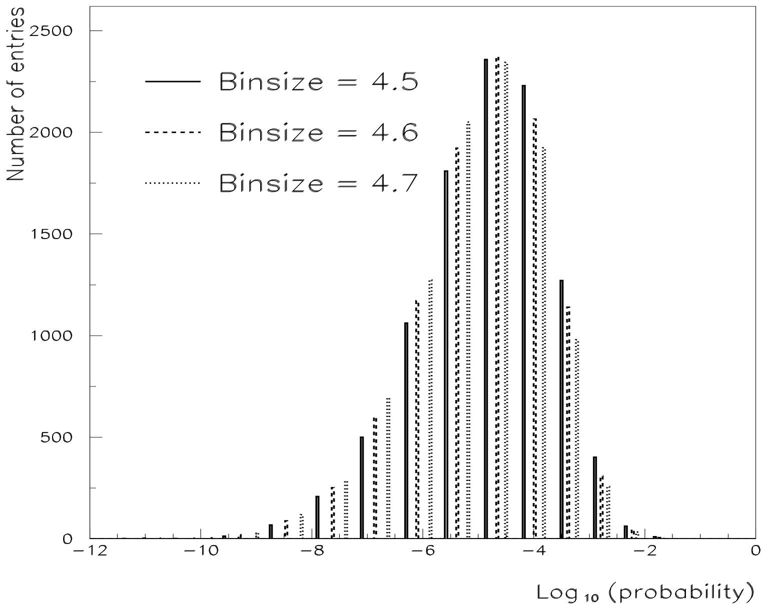

Figure 3: Probability distribution after 10,000 trials for 3 different bin sizes using square bins, an Nexp of 0.5, and a signal significance of 10^{-4}. The peak for all 3 bin sizes occurs when 12 signal events are observed in the bin. The corresponding (peak) probability decreases with increasing bin size because Nexp increases (Nexp = 2.531. 2.645, and 2.761, respectively). This is the reason for the jagged behavior of the maximum probability curve seen in Figure 3.

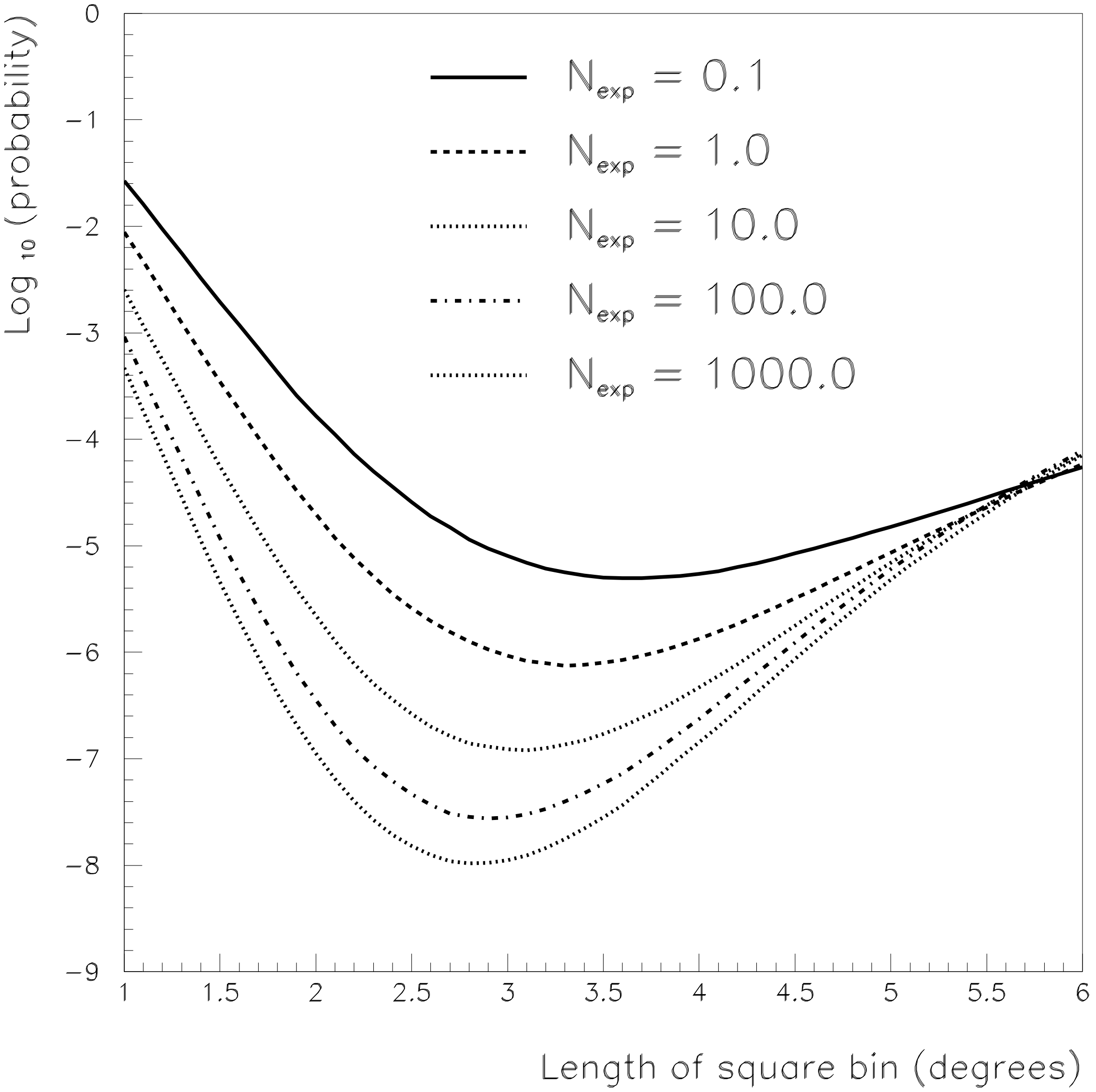

Figure 4: Average of log10(probability) vs. bin size for 10,000 trials using square bins and $\sigma = 1.0^{o}$. The optimal bin size for each Nexp is the bin which give the lowest average log10(probability), which is the minimum for each curve. As would be expected, the optimal bin size increases as Nexp decreases.

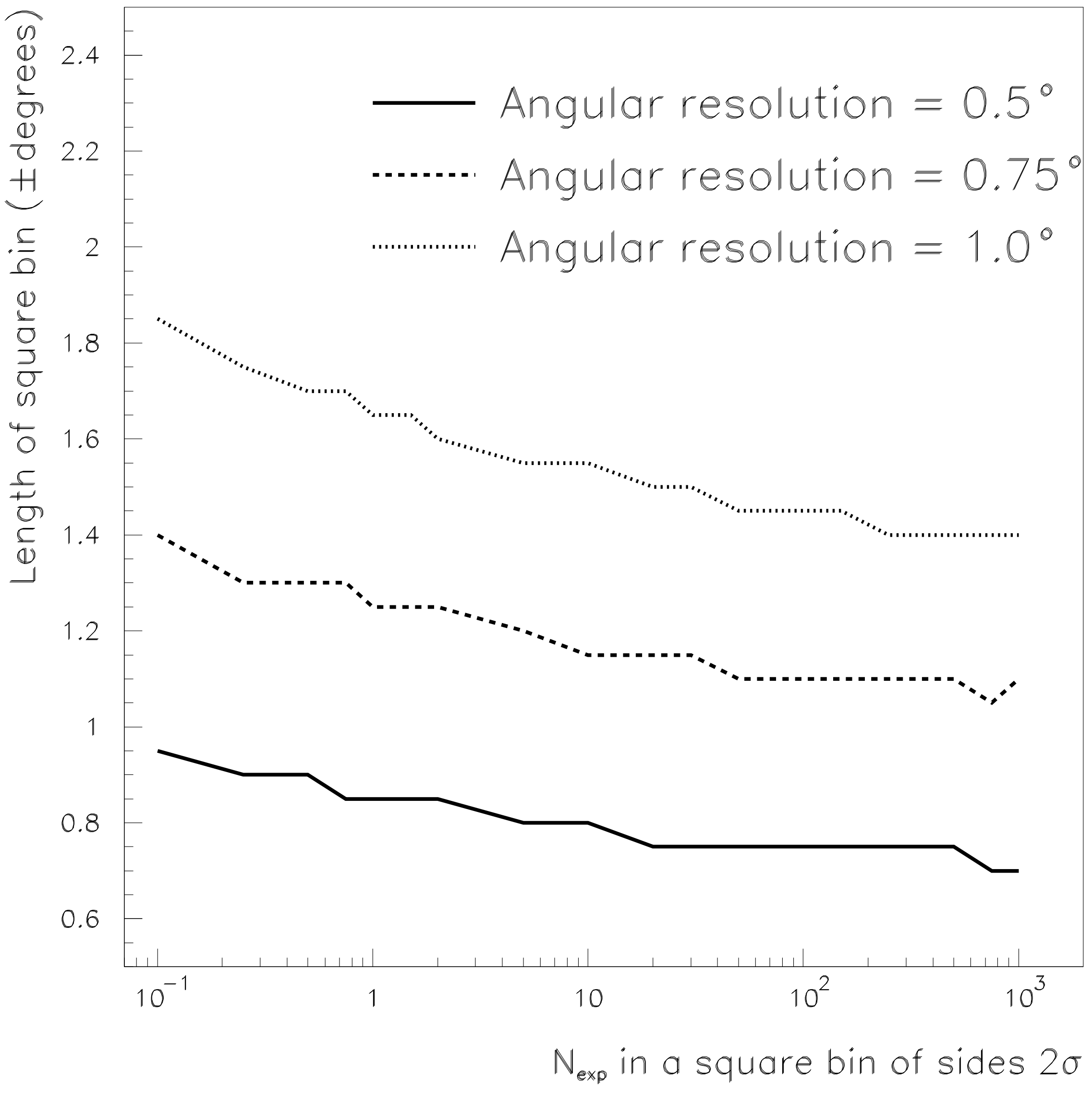

Figure 5: Results of optimal bin size vs. Nexp for 3 different angular resolutions using square bins and a signal significance of 10^{-4}.

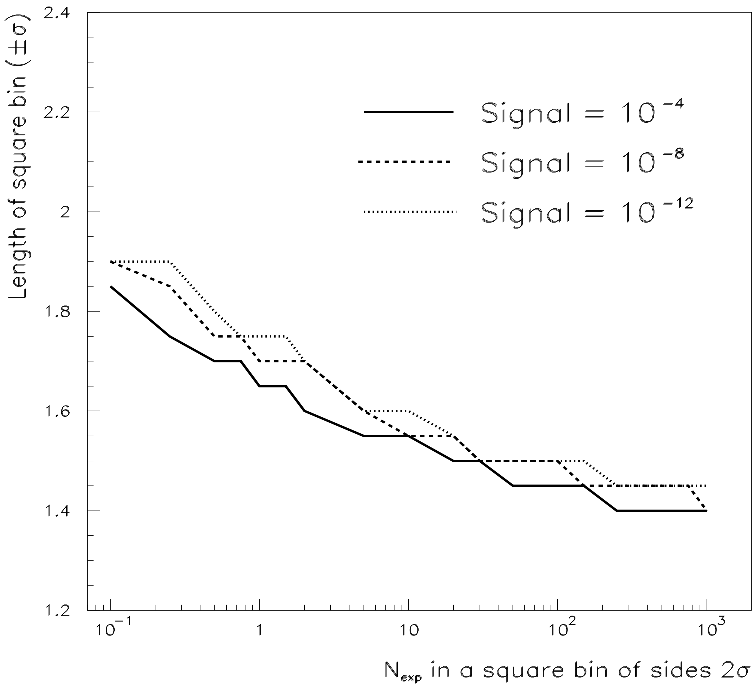

Figure 6: Results of optimal bin size vs. Nexp for 3 different signal strengths using square bins for $\sigma = 1.0^{o}$.

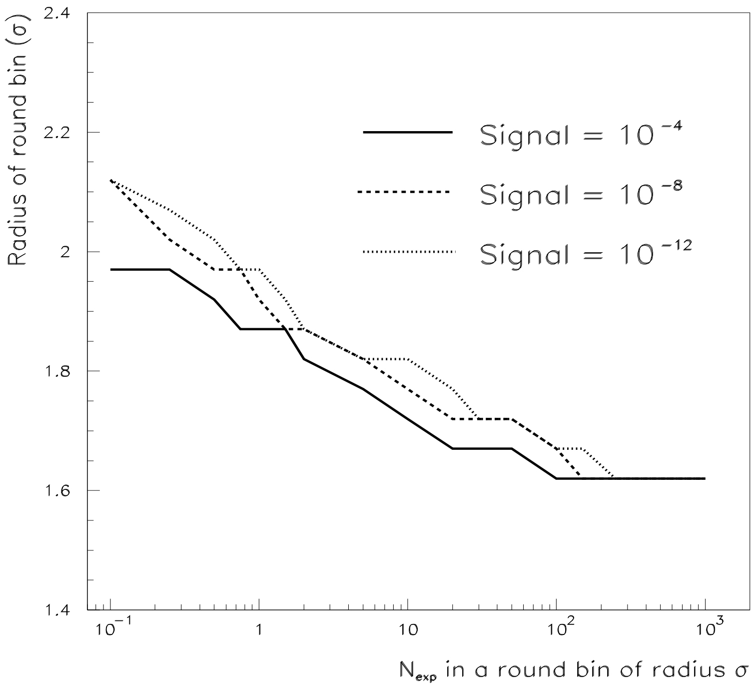

Figure 7: Results of optimal bin size vs. Nexp for 3 different signal strengths using round bins for $\sigma = 1.0^{o}$.

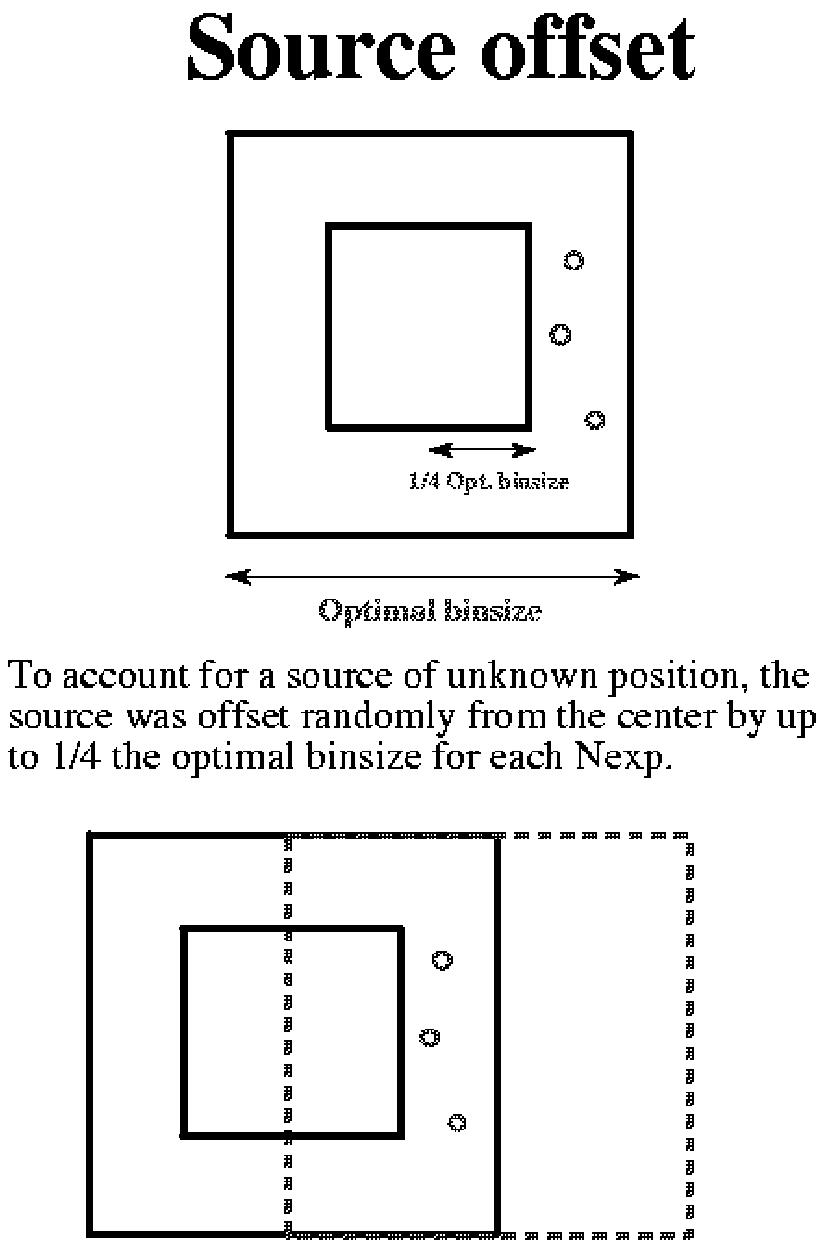

Figure 8: How an unknown source position was simulated. The source center was offset randomly by up to $\pm 1/4$ the optimal bin size for that particular Nexp. Anything offset beyong that (see the 3 open circles) would put the source center close to the center of the adjacent bin when the bins were offset to do the oversampling (assuming oversampling by $1/2$ a bin size).

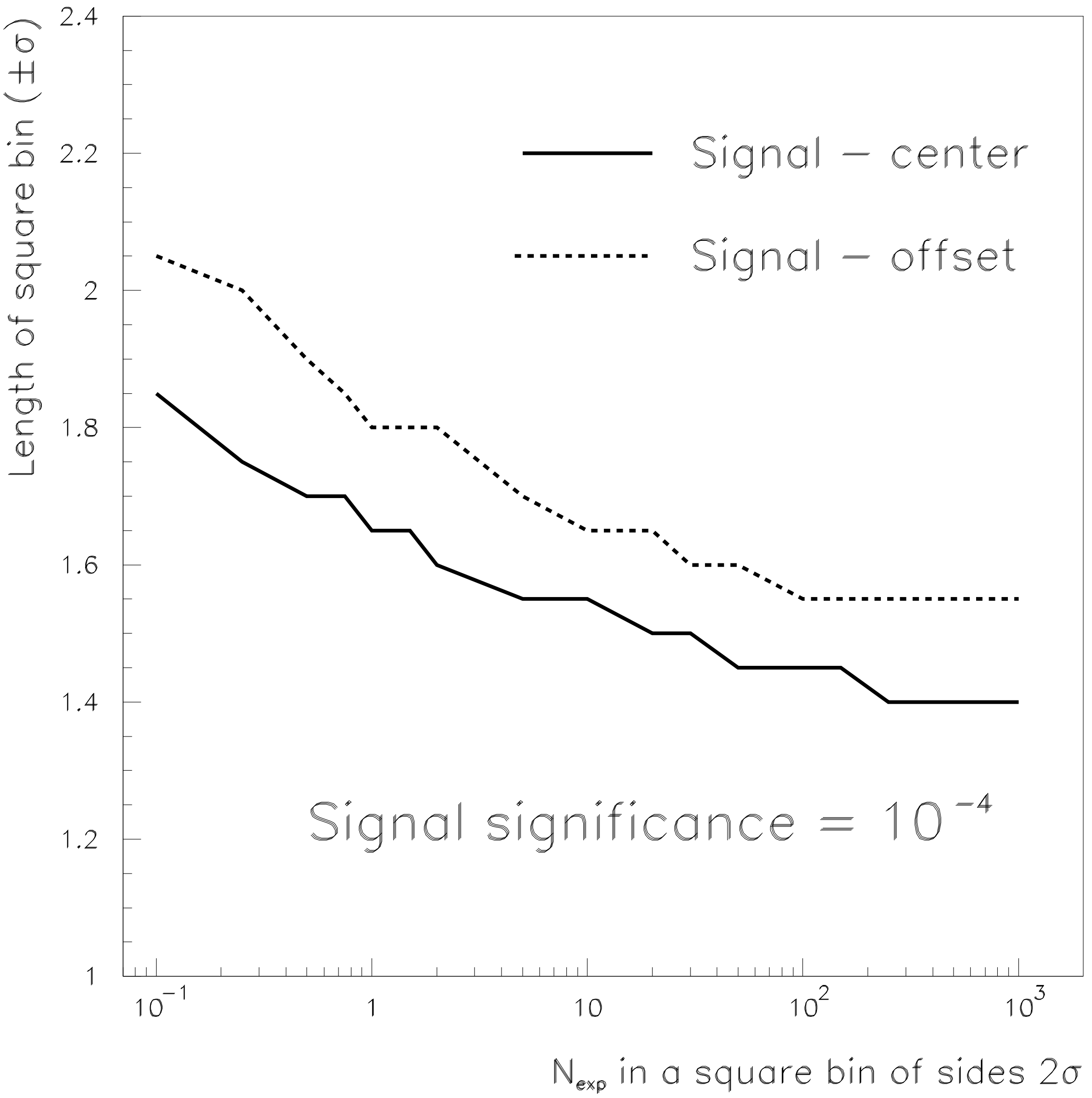

Figure 9: Optimal bin size vs. Nexp for source position known and unknown ($\sigma = 1.0^{\circ}$ and using square bins). The optimal bin size when the source position is unknown is 10-15% larger than if it is known (assuming oversampling by 1/2 a bin).

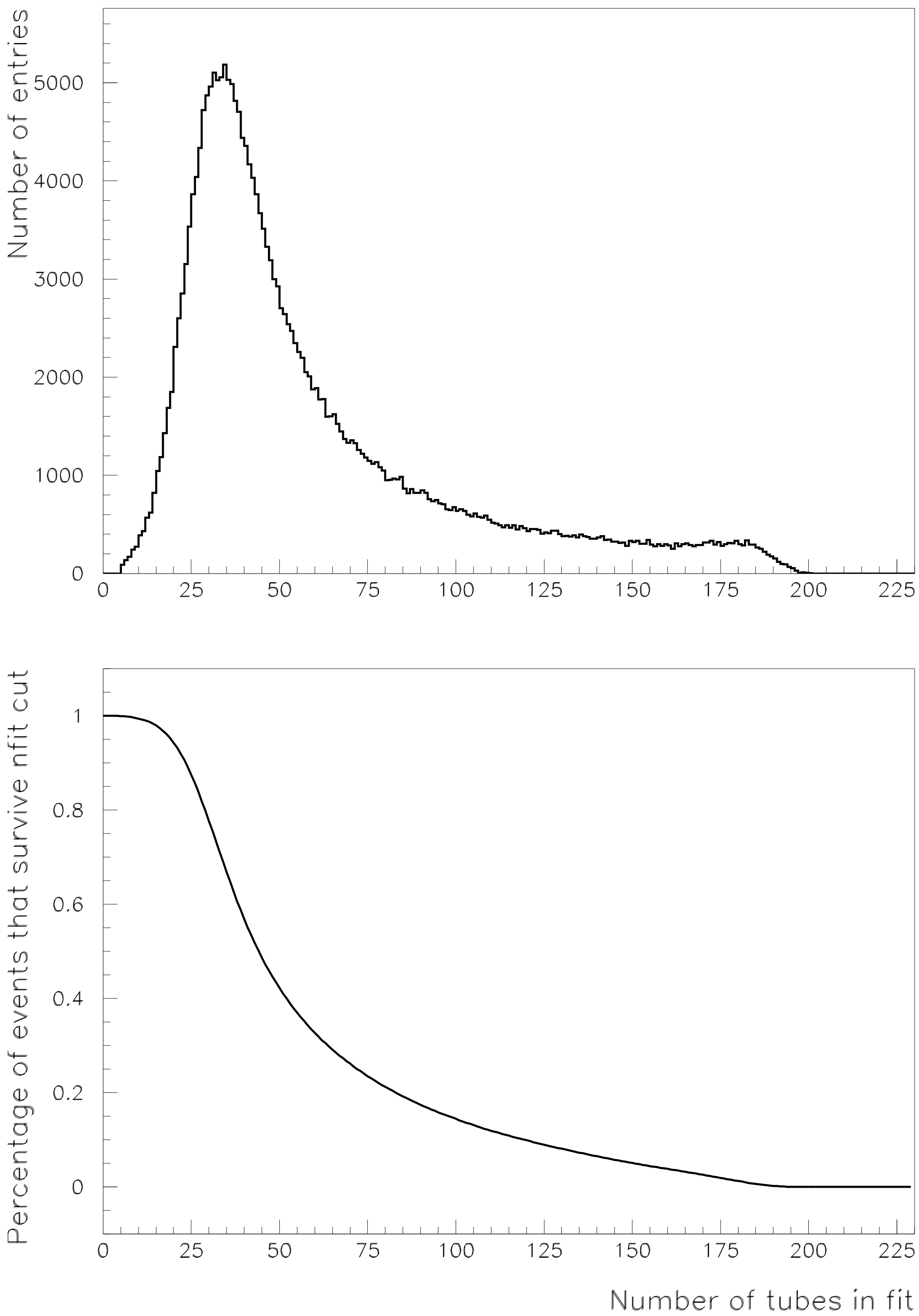

Figure 10: Nfit distribution (top plot) from one subrun of data (MGTPRC_0139_58.dat) and the percentage of events that survive each nfit cut (bottom plot).

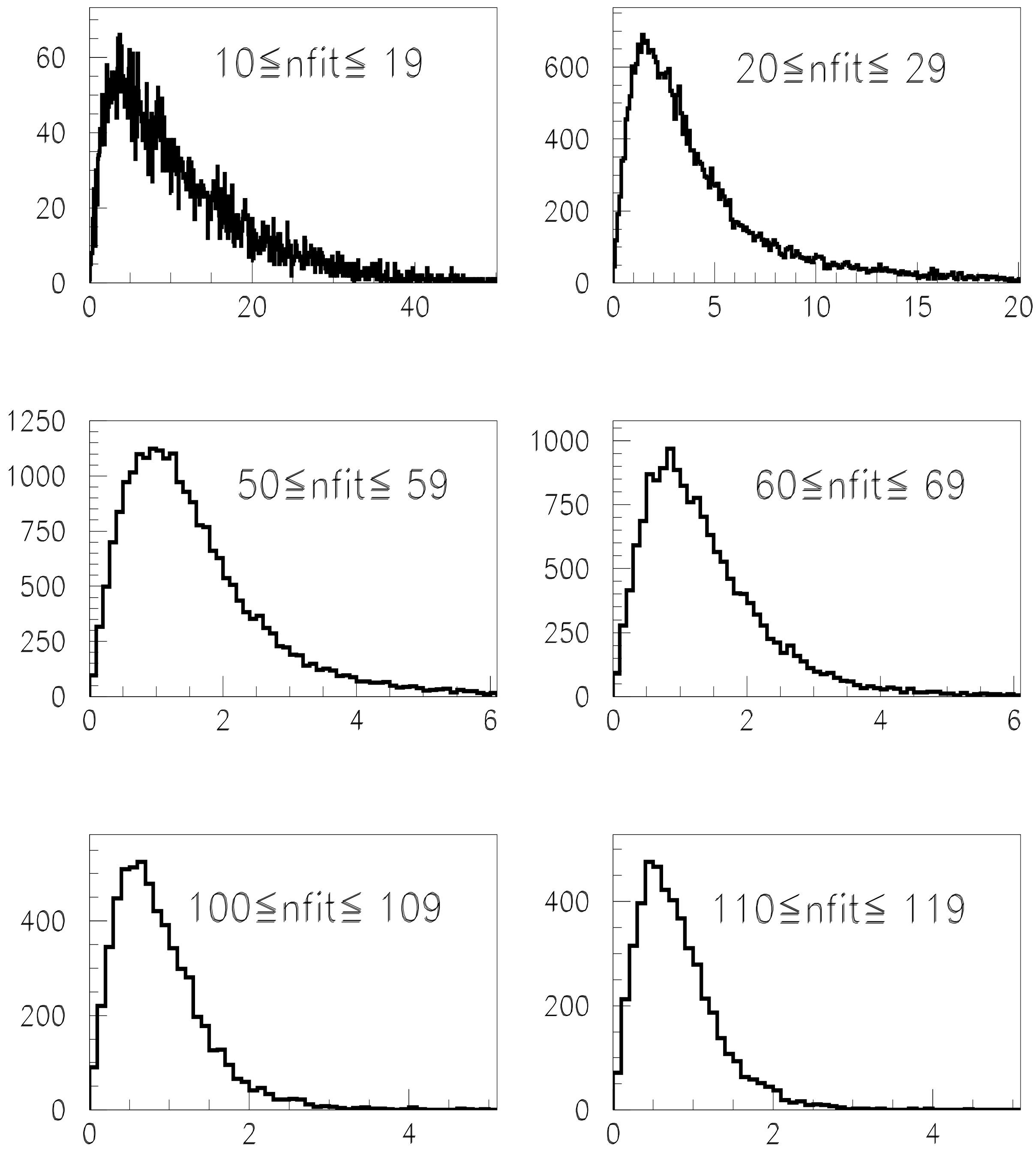

Figure 11: Deleo/2 distributions for 6 different nfit bands. Note that the x-axes have different scales

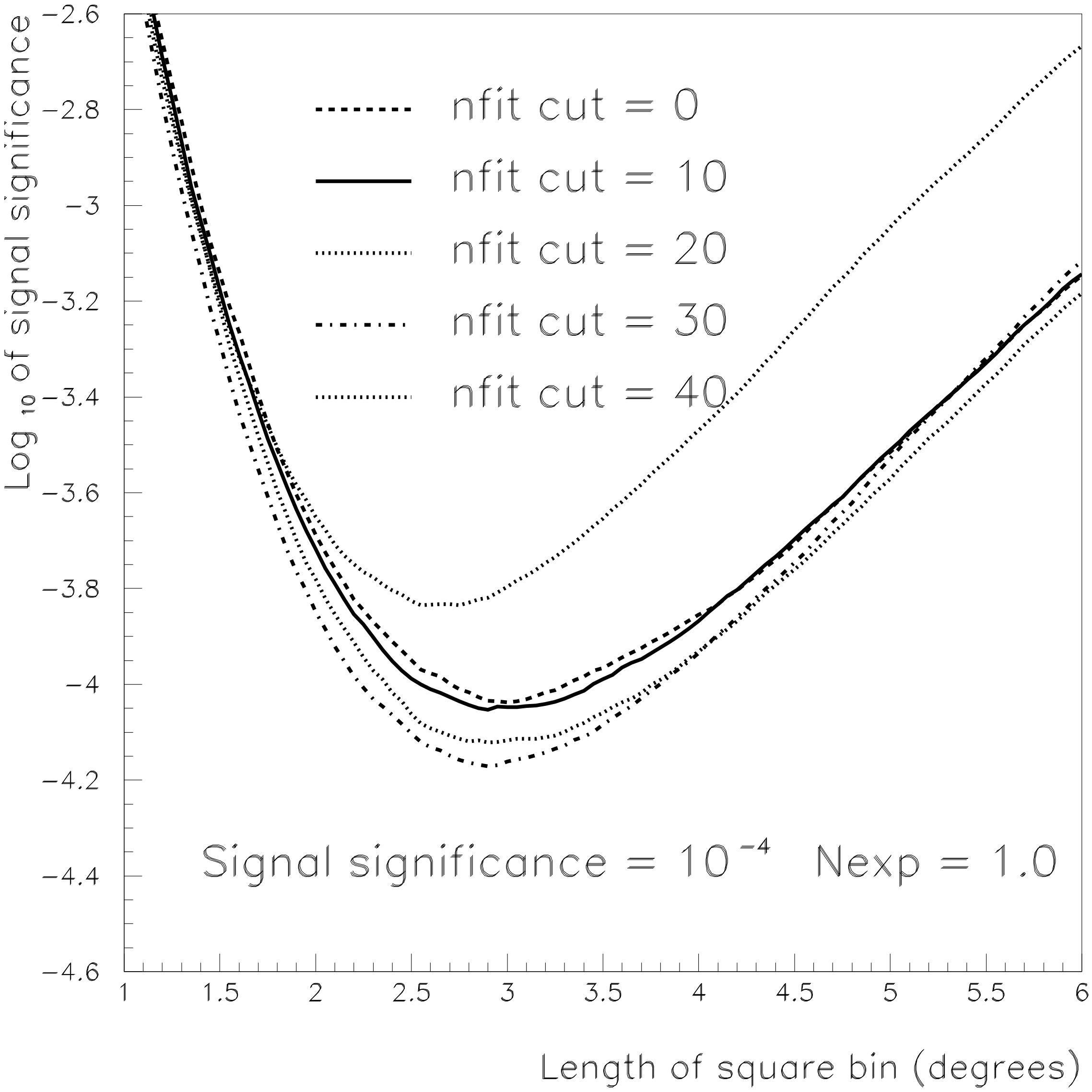

Figure 12: Results after 10,000 trials of the average log10(probability) vs. bin size for a known source position and 5 different nfit cuts. The nfit cut and bin size which has the lowest average log10(probability) is chosen as the optimal nfit cut and bin size. In this case, for an Nexp of 1.0 (in a 2^o square bin), the optimal nfit cut is 30 and the optimal bin size is 2.9^o

	NO OFFSET		OFFSET	
Nexp	**Best Nfit**	**Optimal Binsize**	**Best Nfit**	**Optimal Binsize**
0.10	20	3.60	20	4.1
0.25	20	3.40	20	4.0
0.5	30 (20)	3.05 (3.10)	25 (20)	3.65 (3.55)
0.75	25 (20)	3.00 (3.00)	25 (20)	3.25 (3.35)
1.0	30	2.90	30	3.20
1.5	30	2.75	30	3.30
2.0	25 (30)	2.85 (2.70)	30	3.10
5.0	30	2.55	30	2.85
10.0	30	2.50	30	2.80
20.0	30	2.45	30	2.60
30.0	30	2.40	30	2.65
50.0	30	2.35	30	2.60
100.0	30	2.25	30	2.50
150.0	40	1.90	40	2.15
250.0	40	1.90	40	2.15
500.0	40	1.90	30 (40)	2.30 (2.0)
750.0	40	1.80	40	2.05
1000.0	40	1.85	40	2.15

Figure 13: Results using Milagrito's 'real' angular resolution. For each Nexp, the nfit cut and bin size which gave the lowest average log10(probability) was chosen as optimal. For Nexp = 0.5, 0.75, 2.0, and 500.0, the numbers in paretheses are the nfit cut and bin size used even though they were not the optimal (however, the average log10(probability) was approximately the same - see Figure 14.). This was done so the data could be broken up into three regions, all with the same nfit cut: Nexp ≤ 1.0, $1.0 <$ Nexp ≤ 100, and Nexp > 100.0

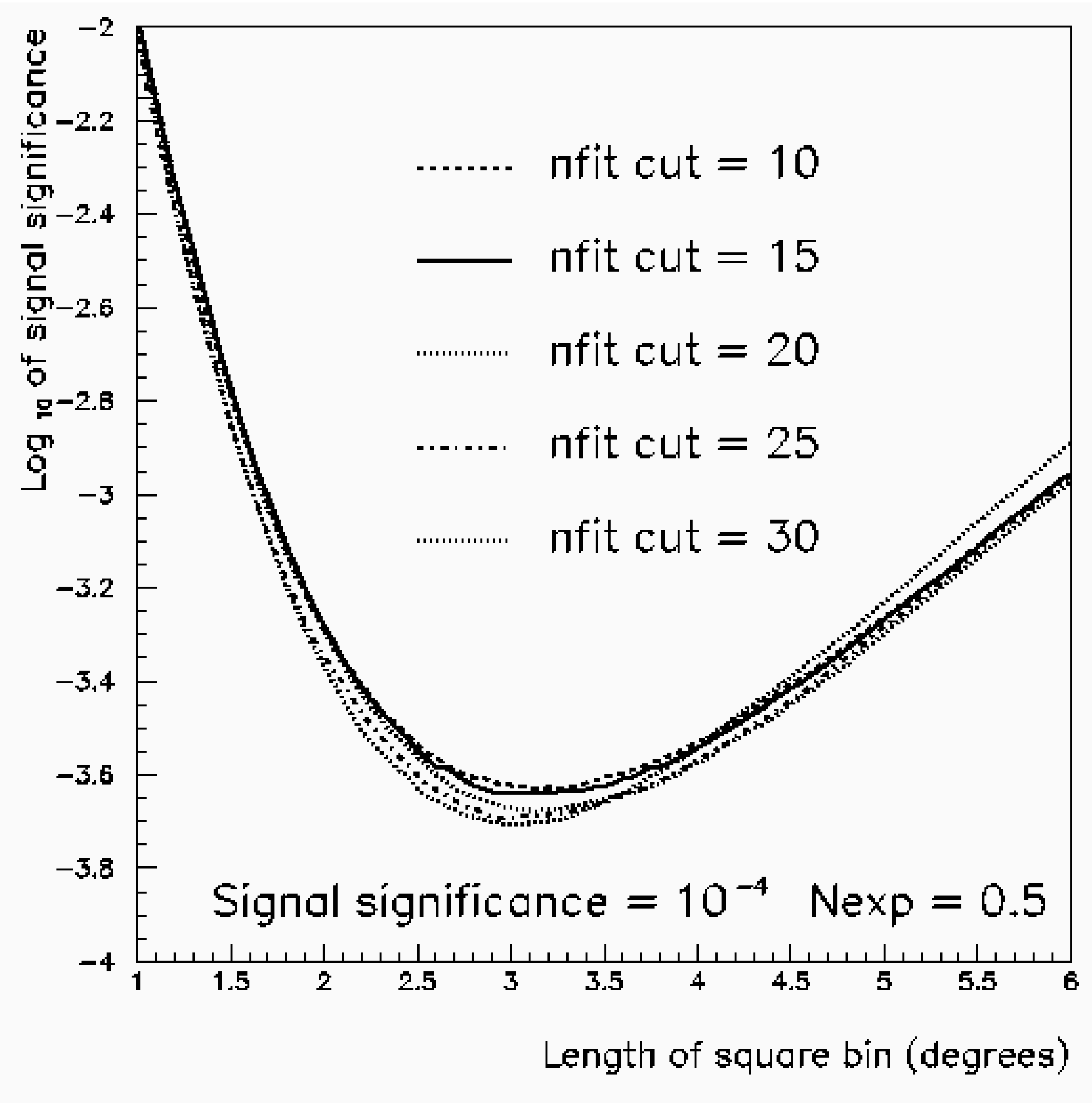

Figure 14: Results after 10,000 trials of the average log10(probability) vs. bin size for a known source position, an Nexp of 0.5, and 5 different nfit cuts. The optimal nfit cut and bin size was 30 and 3.05° which had an average log10(probability) of -3.705. However, the nfit cut and bin size used was 20 and 3.05° which had an average log10(probability) of -3.676. This was done in order to simplify the results (see Figure 13.).

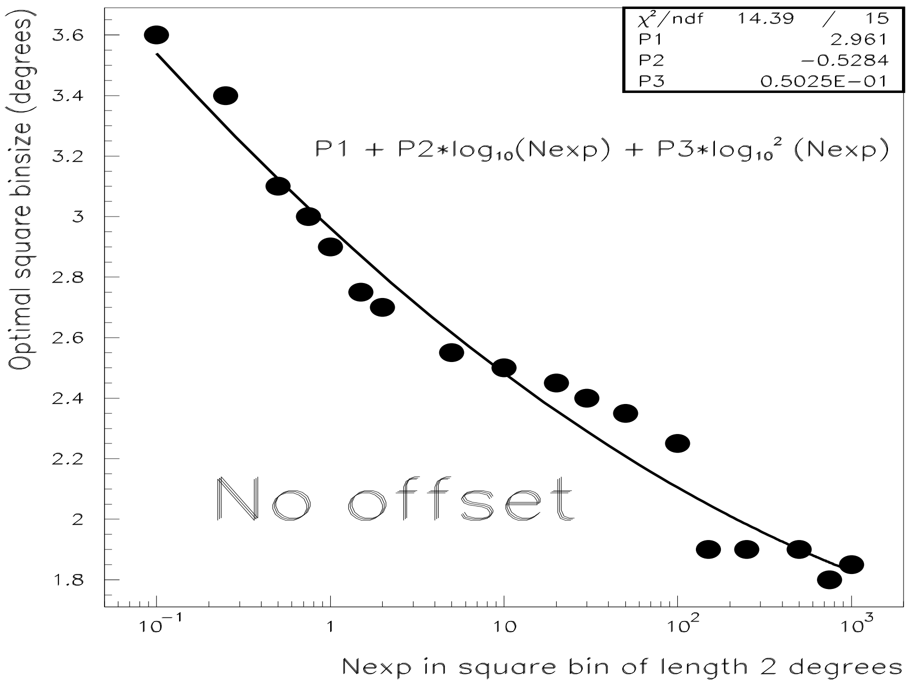

Figure 15: Results of optimal bin size vs. Nexp for a known source position using Milagrito's angular resolution

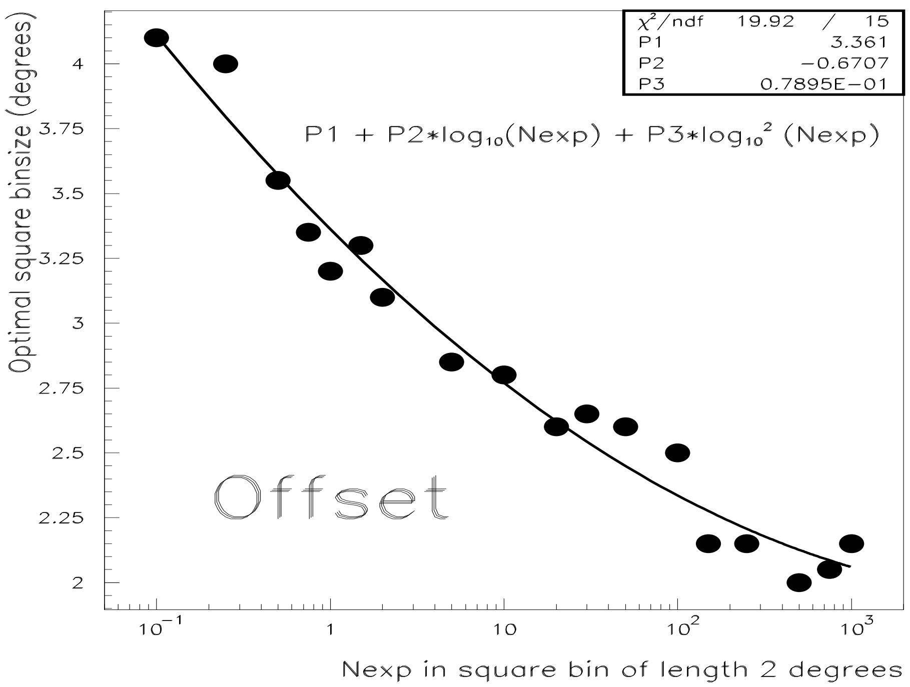

Figure 16: Results of optimal bin size vs. Nexp for an unknown source position (assuming oversampling by 1/2 a bin) using Milagrito's angular resolution

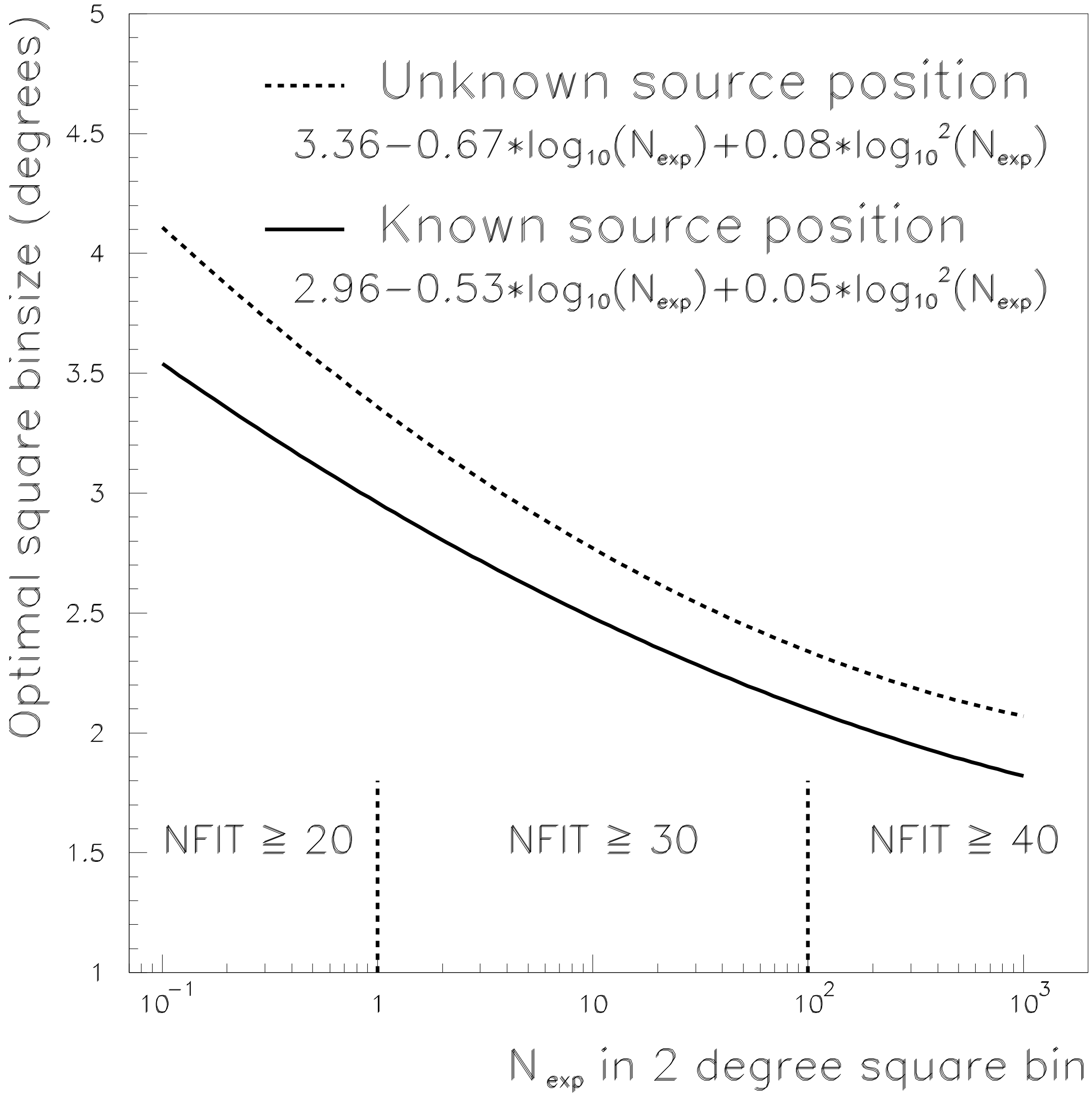

Figure 17: Final results of optimal bin size vs. Nexp for both known and unknown source positions using Milagrito's angular resolution. The data is broken up into three regions: 1) Nexp $<$ 1.0 nfit cut = 20 2) 1.0 Nexp $<$ 100.0 nfit cut = 30 3) Nexp $>$ 100.0 nfit cut =40

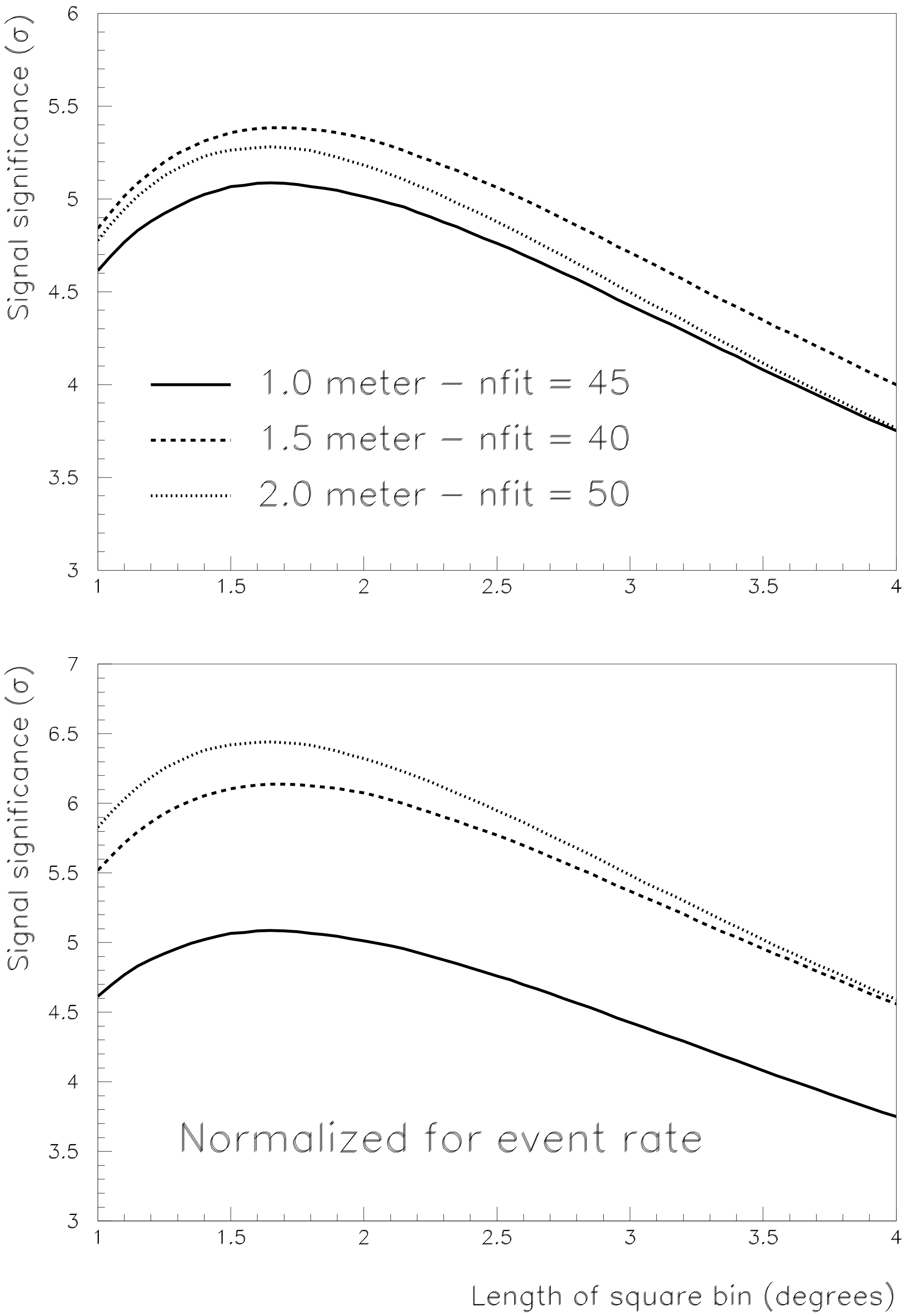

Figure 18: Signal significance vs. bin size for Milagrito at the 3 different water depths. For each water depth, I looked at nfit cuts from 30-60 in steps of 5. Shown in the first plot is the nfit cut (and bin size) which gave the highest signal significance (in σ). In the bottom plot, I corrected for the relative event rates at the different water depths.

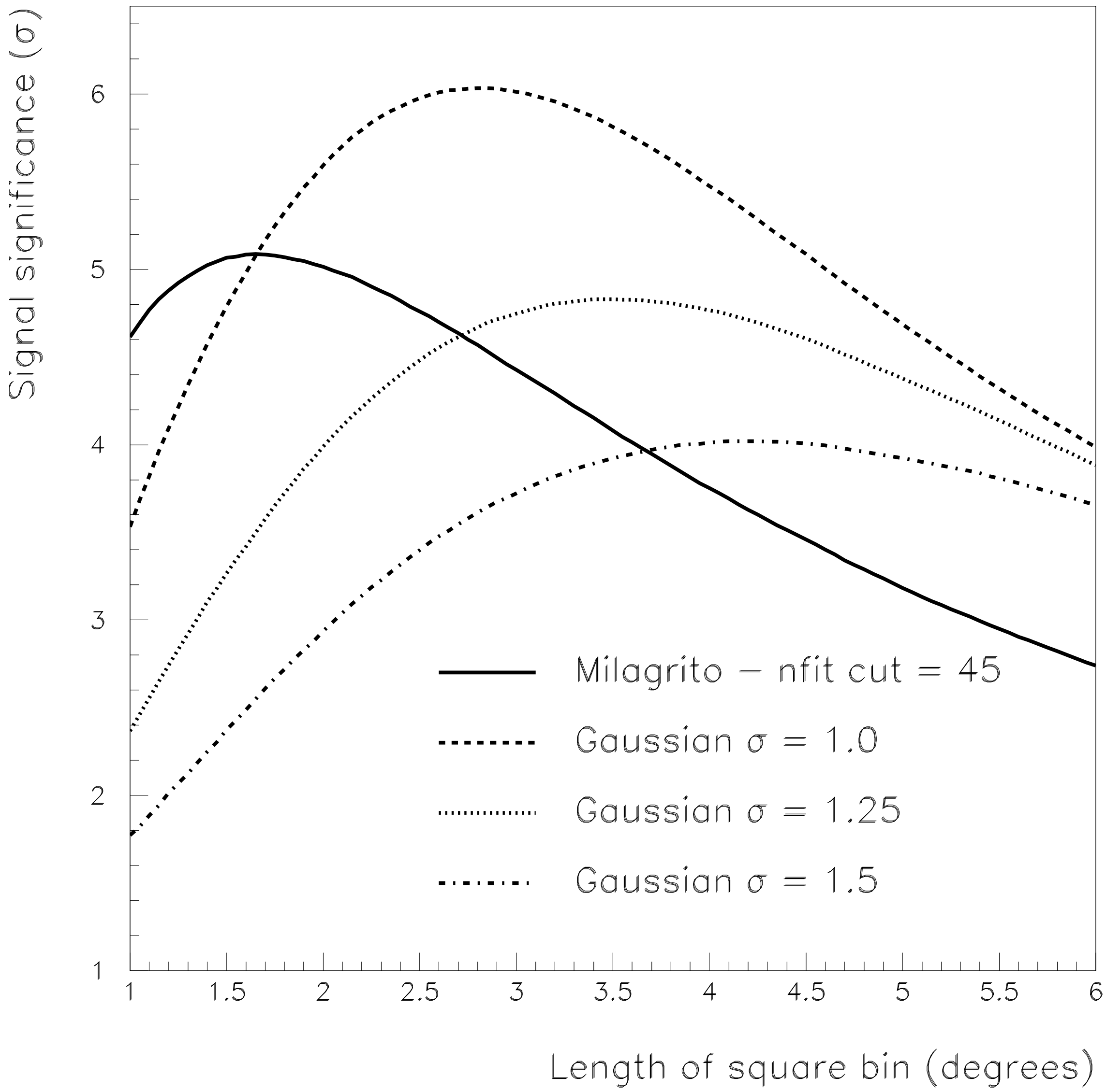

Figure 19: Signal significance vs. bin size for Milagrito (at 1.0m water depth with an nfit cut of 45) and for a detector with Gaussian resolution. If one looks at where the curves peak, the angular resolution of Milagrito with an nfit cut of 45 is comparable to a detector with a resolution of about $\sim 1.20\sigma$.

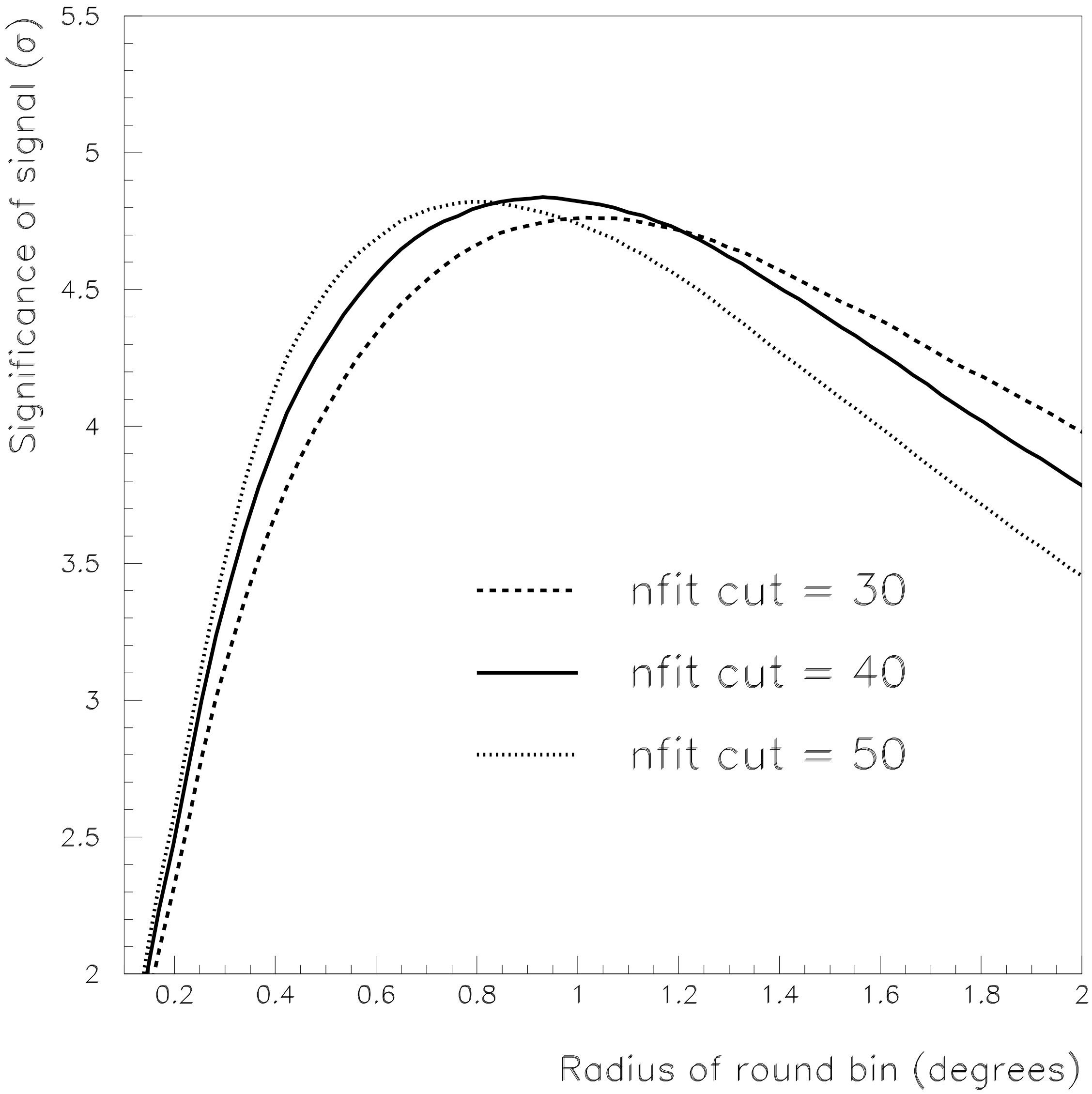

Figure 20: Signal significance vs. nfit cut using round bins for Markarian 501 for 3 different nfit cuts.