PHY340 Data Analysis Feedback
Group P01 doing Problem P4

# Data Analysis

The data analysis is well thought through and thorough, with consideration of several different reconstruction strategies as well as the effects of changing the binning and restricting the range over which the mass peak is fitted. Unfortunately much of this is in fact wrong, as a result of a number of fundamental errors of interpretation.

The first point is that although, as you say, “It is known that the mass distribution...takes the form of a Cauchy curve”, this does *not* mean that you should—even in principle—be *fitting* a Cauchy curve. The reason for this is that there are two contributions to the observed width of a mass peak in an experiment: the natural width of the resonance, which does have a Cauchy form, and the experimental resolution, which is usually approximately Gaussian. The observed peak is in principle a convolution of these two effects, but often one dominates the other. This is definitely the case for the Higgs, whose natural width in the Standard Model is 4 **M**eV—i.e., com­pletely negligible compared to the experimental resolution. There is thus no reason to expect any­thing other than a Gaussian form for the peak (and hence no reason to expect that restricting the fit range will improve the fit).

You do see a reduction in the uncertainty as you restrict your fit range, but this is not because there is an underlying Cauchy distribution. In fact, your quoted errors are simply wrong: the RMS residual is the root-mean-square difference between each data point and the expectation from the fit, and this is *not* a measure of the error in the mean. If your data points were all measurements of the same quantity, the error on the mean would be the RMS residual divided by $\sqrt{N}$, where *N* is the number of points. In fact, your data points are histogram bin contents, not individual mass measurements, so it’s not quite that simple. Your own fit results, however, quote the 95% CL range in the mean values (roughly ±2σ) as (129.1, 132.5) for all the data, and (128.5, 133.2), (128.9, 131.9) and (129.7, 131.8) for progressively restricted samples, so you could use these to quote, respectively, 130.8±0.8, 130.9±1.2, 130.4±0.8 and 130.8±0.5. Although the most restricted sample does give the lowest formal error, this is at least partly because fit­ting a Gaussian over such a restricted range is a procedure of rather questionable validity. (I am actually surprised that you managed to get reasonable results out of your ±0.25σ data: this sample is essentially flat, and it is amazing that Matlab produced a well-constrained Gaussian fit. When I tried to fit a Gaussian over a ±0.5σ range using an online non-linear fit package, the fit was completely unstable.) In general, the removal of outliers is a procedure that should only be adopted when there is a very good reason, which is not present in your case (your unrestricted Gaussian fit shows ***no*** evidence of problems in the wings). It is worth noting, though in fact you do not note it, that the $\overbar{R}^{2}$ values for the restricted fits get steadily poorer: your fit to the full data sample accounts for 95.3% of the variance of your data, your fit to the most restricted data sample for only 64.5%. By restricting your sample, you are just amplifying the statistical noise.

This brings us to the second point: as far as I can tell from your text, you appear to be quoting χ2 as your measure of goodness of fit (this is certainly what is defined at the top of page 7), when you should actually be quoting χ2 per degree of freedom. If this is correct, then the goodness of fit for the 15-bin fits is actually much better than the 4-bin fits (in the former case, there are 12 degrees of freedom; in the latter case, only 1). Indeed, nobody in their right mind would con­si­der making a 3-parameter fit to 4 data points: you have so little freedom that you are almost bound to get a formally good, but physically meaningless, fit.

The final point is that you don’t seem to take the error bars that you do quote seriously: if you were cor­rect in quoting your fitted mass as 130.7±5.8 GeV/*c*2, you would have *absolutely no grounds* to claim that you have a “systematic error”: there is *no meaningful difference* between 130.7±5.8 and 125.09±0.32 (the difference is 5.1±5.8, clearly consistent with zero).

In fact, of course, you *do* have a systematic error, because you have overstated your uncertainty by a factor of 7 or so, so the dis­cussion on page 8 is worthwhile—though, in terms of internal consistency, it isn’t. However, if you are going to discuss systematic errors, discuss the ones that are actually relevant: as you say, you have a clean Monte Carlo sample, so systematics related to back­grounds are not an issue and need not be considered. Also, some of the systematic errors you describe on page 8 do not obviously relate to mass meas­ure­ments: trigger and reconstruction inefficiencies would clearly have a systematic effect on *cross section* mea­surements, but it is not at all obvious that they would have an effect on *mass* measurements; if you are claiming that they do, then you need to justify that claim. (It may be true: if the efficiency of selecting H → τ+τ− candidates were to increase at higher *m*ττ, for exam­ple—maybe because of cuts designed to reduce the Z decay background—then you might expect to reconstruct a slightly high Higgs mass. But if this were the case, you would need to *explain* this: just selecting fewer candidate events does not automatically bias the mass.)


*Figure 1: mass reconstruction using the collinear mass approximation, from Elagin et al. (2011).*

Having discussed some possible systematic errors, albeit not in sufficient detail, you then seem to claim that they are irrelevant (“the systematic errors that would normally compromise the results of a real experiment are not present in this case”). But this is certainly not true of your final potential systematic, namely the fact that the collinear mass approximation is just that—an *approximation*—and in fact the neutrinos from the τ decays will *not* be precisely collinear with the visible decay products (indeed, in decays with only one charged particle and no neutral mesons, which account for 46.8% of all tau decays, the “missing” momentum *cannot* be col­linear with the charged particle, because some transverse momentum is needed to account for the invariant mass of the τ). Your Monte Carlo simulation may be unrealistically background free, but that does not mean that it does not correctly model the τ decay kinematics, so this systema­tic error *will* be present in your sample. It is not intuitively obvious what it does, but Elagin et al. (2011)[[1]](#footnote-1) say that “The collinear approximation is also very sensitive to the *E̸*T resolution and tends to over-estimate the ττ mass,” which is consistent with your result. (This paper is not hard to find: Googling “tau mass reconstruction” brings it up as the first hit.) However, the cause of the high average in Elagin et al. appears to be an asymmetry in the reconstructed mass peak (see figure 1 above), rather than a shift in the peak position, whereas you do seem to see a shift in the peak. On the other hand, Elagin et al. are analysing a sample in which both taus de­cay hadronically (therefore, only one neutrino per tau), whereas your sample has one tau de­caying leptonically (two neutrinos), and this difference will certainly affect the kinematics (in particular, for a single neutrino *E*ν = |**p**ν|, whereas for two neutrinos *E*νν ≠ |**p**νν|, as the two neu­trinos will not in general be travelling parallel to each other).

Perhaps this is a consequence of group members having different types of coding experience, but you do seem to have taken a needlessly convoluted approach to your data analysis: selecting data using Python, binning in Excel and fitting in Matlab is, frankly, simply silly (Python is en­tirely capable of binning data, fitting data and plotting data, so once you have read your data into Python there is no point in not doing the rest of the analysis in Python too).

In summary, it is quite difficult to mark your data analysis: your approach is excellent, but the actual conclusions are not well justified, because of fundamental flaws in your assumptions. The two markers took different views of this, and the final average combines two somewhat different marks.

Average mark for this section: 39.5/50

# Data Presentation

Your data presentation is clear and comprehensive. Numerical values in the text are quoted to reasonable precision (one might argue that they are a bit overprecise given your (wrong) er­rors: 130.4±5.8 could be rounded to 130±6, for example). You have plotted the right quantities, and your figures have numbers and captions. However, the plots themselves would be better displayed as points with statistical error bars rather than histograms, and you need to *explain your fit*: just saying that you used “the Matlab curve fitting app” is *not* helpful (your reader pro­bably doesn’t care what tool you used). Was this a least-squares fit or a maximum likelihood? If it was a least-squares, was it weighted or unweighted? What minimisation algorithm was used? These are important points, which your reader *does* need to know. Given that you quote various measures of goodness of fit (SSE, *R*2, adjusted *R*2, RMSE), you need to explain what these are: although there are incomplete definitions in appendix B, these are not referred to in the text or the figure captions. (The definitions are “incomplete” because you do not define what you mean by “sum of squared error” or “sum of squared regression”: certainly the “sum of squared error” is *not* the sum of the squared errors, since if it were it would be equal to the number of points in the sample (the Poisson error on the content of a histogram bin is the square root of the contents, so the square of this is just the number of entries in the bin, and hence the sum of these should be the number of points in the sample).) It is disconcerting that one of your fit outputs is misleadingly labelled: the mode *should* be the value that appears most often in the data set, i.e. the *x* value corresponding to the maximum value of *y*, whereas the variable labelled “mode” in your data is *y*max itself. This *is* related to one of the parameters of a Gaussian fit, but it is *not* the mode of the distribution and should not be labelled as such.

Average mark for this section: 21.35/30

# Style

The report is presented in an appropriate style, with a decent abstract, suitable section headings and an awareness of standard scientific style. It appears to have been proof-read: there were few if any typos. There are, however, some small but noteworthy issues:

* Do not put definitions of symbols in an appendix! Appendices are, by definition, things you do not need to read in order to make sense of the paper: derivations of equations belong there, definitions of symbols most certainly do not.
* Number your principal equations, so that they can be referred to later in the text.
* Equations in text generally form part of a sentence, which continues below the equa­tion: in particular, the definitions that follow the equation (“where *x* and *y* are...”) are part of the same sentence and do *not* get a capital letter.
* χ2 has been written as X2 in the text, for some reason (maybe this counts as a typo, but it happens consistently: it’s not a one-off). Note that even in san-serif fonts, χ is *not* the same as X: the former is a lower-case letter with descenders that drop below the level of the line; the latter is an upper-case letter that rises above the level of standard lower-case letters.

These are not major issues, and overall the report looks nice and reads well, but they are things that you should fix.

Average mark for this section: 15/20

Overall average mark: 75.85%.

1. A Elagin et al., *Nucl. Instrum. Meth.* **A654** (2011) 481-489. [↑](#footnote-ref-1)