Report on JINST_042P_1214

Date: January 22, 2015

Author(s): P. D. Dauncey, M. Kenzie, N. Wardle and G. J. Davies

Title: Handling uncertainties in background shapes: the discrete profiling me

Received: 2014-12-25 20:59:59.0

Referee report

The paper provides a detailed description the new background modeling technique used in the final Run1 CMS H->gamma gamma analysis. In this method, the data is allowed to choose between several possible functional forms to describe the background. The type of function is described by a discrete parameter in the statistical model, allowing to estimate an uncertainty on the modeling using the usual profile likelihood technique, albeit in a discrete context.

The method is innovative, and represents an interesting alternative to the ones used previously for the H->gamma gamma analyses of ATLAS and CMS. It is also potentially applicable to other analyses using similar techniques (e.g. the H->mumu analyses). Due to the high profile of these analyses, it presents a high scientific interest. It is well-written and generally pedagogical and clear. An overview of the method was already presented in the CMS H->gamma gamma paper (EPJ C 74 (2014) 3076) but in much less detail, so in my opinion this does not detract from the novelty of the present paper

On the other hand, the bulk of the present study is limited to the specific context of the CMS H->gamma gamma analysis, and there is no discussion

on whether its results are more widely applicable. The paper would benefit form a somewhat broader scope, in particular more discussion of whether the method can be expected to be valid in other contexts. Specific points that could be considered are 1) the domain of validity of the asymptotic formulas used to derive the uncertainties and 2) the choice of the set of functions used. These points are detailed below. Other points are also listed in items 3) to 8).

I believe the work merits publication, after these issues are addressed. Some items, especially 1) and 2), would require a significant amount of work to address in detail, and I do not insist this be done – however including at least a discussion of these points would be worthwhile in my view.

===

1) Usage of asymptotic formulas

The modeling uncertainty is obtained from a profile likelihood technique using asymptotic approximations. This is an important property, making the method easy to apply, and the main result of the paper is to show that good coverage is obtained usimethodng this . However this is shown only in the context of the CMS H->gamma gamma analysis, and it is not clear how widely this is valid.

One would naively expect that the distribution of the profile likelihood is not purely a chi2, but rather the sum of several terms:
- a chi2 term corresponding to the case where the functional form chosen by the fit is the true one
- other terms for the cases where another function than the true one was chosen by the fit. A priori these would be described by non-central chi2 distributions, which would break the approximation used in the paper.

The results in the paper seem to show that this is negligible in the CMS H->gamma gamma case, either because the probabilities to chose the "wrong" function are small, or because the non-centrality parameters are small.

This could be studied for example by looking at the distributions of the profile likelihood ratio in toys, separating the fits to each of the allowed function types. This, along with a wider discussion of the applicability of the

method, would serve both to better justify the good properties of the method for the CMS H->gamma gamma case, and to provide tools which could be used to study its application to other cases. (the plot on Fig. 14 gives some information along these lines, but is less straightforward to interpret; it is also given without the separation of the fitted function types).

===

2) Definition of the set of functional forms

One of the key ingredients of the method is the set of functional forms that are considered. The analysis uses a reasonable set, but it would be helpful to discuss the effect of this limitation, for instance the fact that polynomials are only considered up to order 5 (this point is briefly mentioned at the top of p.20, but the conclusion is unclear). This is relevant to some results in the paper: for instance the lower part of the plot in Fig. 11 would likely show a larger bias if higher-order polynomials were considered, which may have an impact on the conclusion at the bottom of p.13.

The uncertainty derived by this method is also accurate only if a sufficiently large number of functions are used, so that the envelope is correctly sampled. Since the set of functions used is limited, one should probably consider an additional uncertainty due to those functions that would have contributed, but were not considered. This is especially true for low values of c, for which functions with a quite large number of parameters can contribute to the envelope and increase the uncertainty.

This could be checked for instance by including higher-order polynomials in the allowed set, and checking the stability of the intervals obtained. It is probable that this would show that in the case of the CMS H->gamma gamma analysis, this additional uncertainty is negligible. However it could be an interesting check, and also be useful for other applications of this method.

===

3) I did not understand the meaning of "optimal" in the middle paragraph of p.20. From the discussion on p.15, it seems c should be adjusted so that the coverage is accurate (i.e. favoring the p-value methods over Akaike). How-

ever here the point seems to be to *minimize* the uncertainty, which seems contradictory. Sorry if I am missing something obvious

4) The plots on Figs. 6 and 13 show increasingly good agreement between the toys-based and asymptotics-based coverage, reaching per-mil-level agreement for the 99.7% interval. I think this may be partly due to the fact that both numbers are by definition <1, so that there is little space for differences. Especially for these high CL intervals, one could consider showing the ratio of the size of the intervals for a given CL, instead of the CL levels for a fixed interval size. So one would for instance compute the 99.7% intervals using both toys and asymptotics, and show the ratio (mu_{hi,asympt} - mu_{lo,asympt})/(mu_{hi, toys} - mu_{[lo,toys}). The agreement would probably be less sensitive to the CL level used, and the test also arguably more relevant, since the size of the interval is in fact (twice) the uncertainty that one is trying to estimate.

5) Section 5 could be shortened considerably. While it is interesting to note that the method does not scale well in the case of many signal regions, some of the material presented here seem rather specific to the CMS H->gamma gamma analysis and does not need to be so detailed in this paper.

6) Several features of Fig. 12 (pull evolution in the top panel, and a similar feature for high mu in the bottom panel) are not explained in the text. This is probably the same issue as that explained on p.19 in another context (inability of lower-order functions to accurately describe the 5th-order polynomial distribution), but it would be helpful to have this explanation here as well.

7) Equation at the bottom of p.8 : I find the Λ=χ'2 equality somewhat confusing, as Λ is not necessarily exactly a chi2. The same equation without the intermediate =χ'2 part seems clearer. Also, an equation relating χ2, its p-value and χ'2 may be helpful – or at least an explicit reminder that χ2 uses (n_bins - n_par) degrees of freedom, vs. n_bins for χ'2.

8) It may be interesting to the reader to be given the best-fit values of the background parameters for the various function forms, to accompany the signal parameter values already given in Section 3.2.

9) Fig. 8 : since no correction is applied here, the "Λ+correction" labels on the Y axes are slightly misleading.