Open statistical models: CMS view

Andrea Carlo Marini

On behalf of the CMS Stat. Comm.

Introduction

- Why talking about open likelihoods?

(Re)interpretation of the LHC results for new physics

The CMS Collaboration uses likelihood functions for most of the results and searches.
Most commonly the likelihood is in the form:

$$\mathscr{L}(\mu, \overrightarrow{\theta}) = \mathscr{L}_{\text{stat}}(\mu \cdot S + B, \theta)\pi(\theta_0 \,|\, \theta)$$

Although,

- Open statistical models is a better terminology of what we want.
- Likely easier to achieve

Caveat: some points are a personal view not necessarily shared by the all Collaborations

What information we need to have

Nowadays analyses are **A LOT** more complicated:

- Shape based

- Several bins

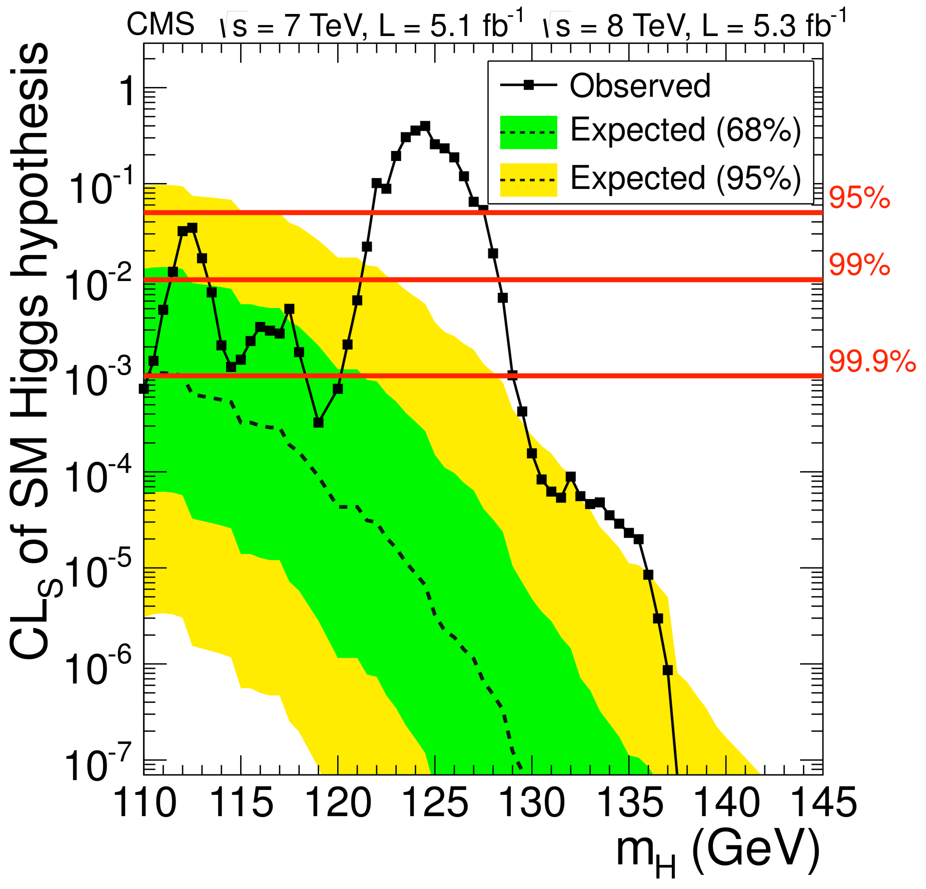
- Several nuisances

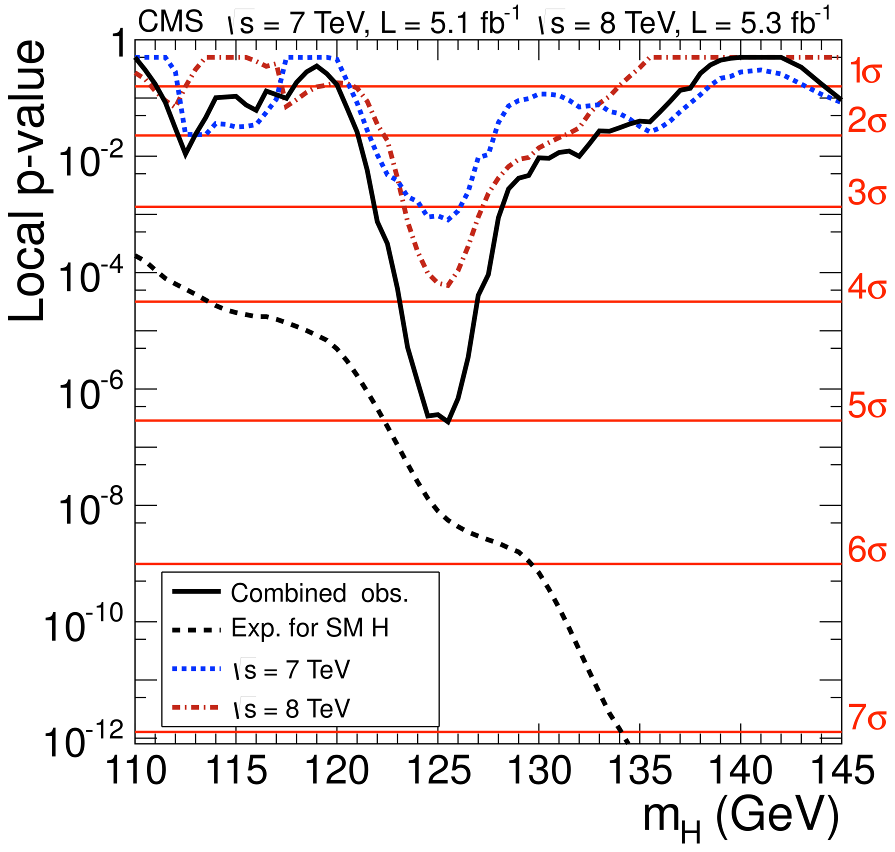
The first step could be to publish the covariance matrix at the best fit value!

Often, it is limited to the POIs
We give a qualitative description of the uncertainties in the papers

Usage of likelihood functions in CMS

- Most of the CMS measurements are performed with likelihood functions
- These are inputs for statistical inference and interpretation. Most common
  - Likelihood ratio confidence intervals
  - P-values of Hypothesis testing (Significance)
  - Upper limits using the CLs criterion





Example of likelihood functions

Definitions: $\mathscr{P}(n\,|\,\lambda) = \dfrac{e^{-\lambda}\lambda^{n}}{n!}$ $\quad\mathscr{N}(x\,|\,\mu,\sigma) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Let's take the log. This is a 'template' binned likelihood (simplified)

r → parameter of interest, θ nuisance parameters.

Two bins, two nuisances :

} Parameters. $\mathscr{L}$ is usually maximised over them

$$\log \mathscr{L}(r,\theta) = \underbrace{\log \mathscr{P}(d_1\,|\,rS_1k_1^{\theta_1} + B_1k_2^{\theta_1}) + \log \mathscr{P}(d_2\,|\,rS_2k_3^{\theta_1} + B_2k_4^{\theta_2}k_5^{\theta_1})}_{\text{Data model}} + \underbrace{\log \mathscr{N}(0\,|\,\theta_1,1) + \log \mathscr{N}(0\,|\,\theta_2,1)}_{\text{Nuisance constraints External measurements}}$$

k_i →Values of the effect of the nuisance parameters.

In the example above are logNormal likelihood

d_i → values of the observation

S_i, B_i → values of the predictions for signal and background.

} Actual numbers In the final evaluation

How will these models be used?

What we think a purpose of publishing them is:

1. **Reinterpretation**

Change the parameter of interests (r) and how they map to the different Signals processes in the analysis

$$\mathscr{L}(r,\theta) \rightarrow \mathscr{L}(r_1, r_2, \ldots, \theta)$$

2. **Combinations**

Take two likelihood functions and map the corresponding correlation of nuisance parameters

$$\mathscr{L}_1(r,\eta), \mathscr{L}_2(r,\nu) \rightarrow \mathscr{L}(r,\theta) = \mathscr{L}_1(r,\eta(\theta)) + \mathscr{L}_2(r,\nu(\theta))$$

3. **Change dataset** to run, e.g., Asimov expected results or pseudo-experiments

The full statistical model

To be able we are actually going to publish the full statistical model:

- A parametric description of the probability density functions (p.d.f.) or a way to construct it from the inputs (see next talk)
- The possibility to exchange the observations with other datasets
  - Asimov dataset
  - Pseudo-experiments
- The list of nuisance parameters and external measurement values and their p.d.f. / constrain type.
  - to derive frequentists toy datasets
- The list of observables (for binned and unbind likelihood models)
- The code for the custom pdf used.

- If we follow this path, it is likely we are going to use the inputs (datacards) we use in our tools.

What could we release?

- We need to start somewhere …
  - 1) It should be small
  - 2) It should be useful
  - 3) It should not be an excessive workload for review and analyser.

3) implies that the physics interpretation of the parameters may not be fully documented.

It is very unlikely that we will release the full statistical model for all the analysis

When?

The CMS Collaboration is using the statistical models for different purposes:

- Single publication
- Reinterpretations
- Combinations

The timeline to publish full statistical models will likely be delayed in time (similar to the open-data)

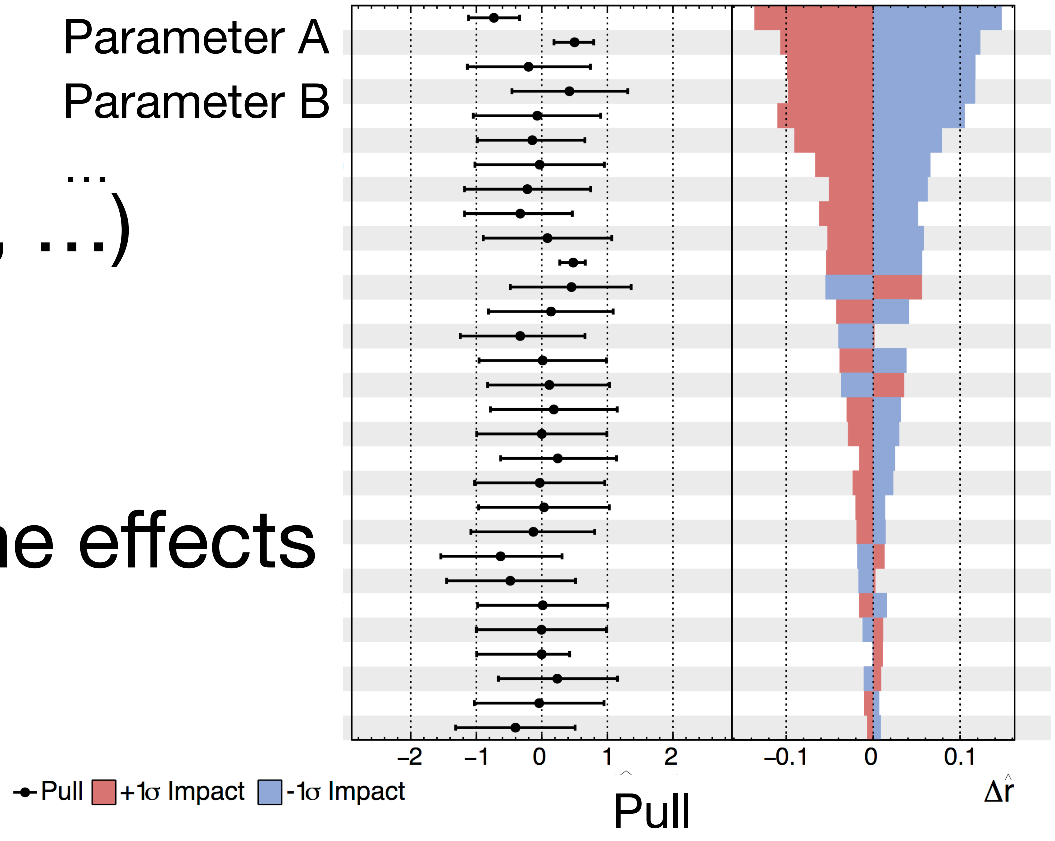
A **caveat** will be necessary to the results derived from as not been either endorsed or review by the collaboration.

A word on pulls & constraints

The usage of Profile Likelihood “constraints” some uncertainties

These nuisances can be:

- background modelling (th. uncertainties)
- Experimental uncertainties (e.g., jes, b-tag, …)



Could be seen as ‘in situ’ constraints of some effects

Sometimes effects are still negligible in the particular analysis

However, when running combinations or reinterpretations,
care should be put in understanding what are these constraints,
If they are now important, and
what do they propagate to.

Interplay with current tools

HEPDATA

Could be used as storage.

We are already publishing there a lot of the digitisation of the current results, and covariance matrixes.

RIVET routines

Publishing rivet routines could still give an approximate way to calculate some new signals.

- although detector effect will not be fully accounted, and could be approximated by the old model or by other means (e.g. delphes). :(

Summary

- The public release of full statistical models is an interesting topic that the nowadays technology enables

- It will enable and extend the life of the analysis results. It’s a possible legacy of the LHC

- There are some limitations in the review and analysis we will apply on

- The technicalities needs to be finalised:
  we are working towards a ‘Combine’ tool public note