PROCEEDINGS OF SCIENCE

Publishing full statistical models of CMS physics analyses

Sezen Sekmen
for the CMS Collaboration[a,*]

[a] *Kyungpook National University, Department of Physics*
80 Daehak-ro buk-gu, Daegu 41566, Republic of Korea

E-mail: ssekmen@cern.ch

The CMS Collaboration has recently approved the public release of full statistical models for its physics analyses, including the data necessary to construct complete likelihoods. The statistical inference tool, COMBINE, essential for this process, is now available under an open source license. This report highlights key features of COMBINE and discusses the publication and utilization of statistical models, including the first released model used in the discovery of the Higgs boson.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

*Speaker

PoS(ICHEP2024)988

1. Introduction

The LHC experiments have established a remarkable physics legacy through hundreds of physics analyses performed on their data. However, maximizing the impact of this legacy relies heavily on ensuring the accessibility of information that will facilitate the future reuse of these analyses, both within the experiments and by the broader high-energy physics (HEP) community. Among the various types of information, the statistical model holds particular importance, as detailed in the community reports [1, 2]. In response, the LHC experiments have intensified their efforts to make these statistical models publicly available.

A statistical model is the mathematical framework used to describe and infer the underlying processes generating observed data. It defines the probabilistic relationship between the observed quantities (i.e., the data) and the model's parameters. These parameters include parameters of interest (POI), such as signal strength or resonance mass, and nuisance parameters, which are not directly of interest but are necessary, e.g., to model various uncertainties of theoretical and experimental origin, such as detector effects, background estimations, luminosity calibration, or cross-section calculations. The value of the statistical model for a fixed set of data, as a function of its parameters, is known as the likelihood. The statistical model provides a complete mathematical description of an analysis and forms the foundation for its interpretation. Publishing statistical models maximizes the scientific impact of an analysis by preserving and documenting its mathematical framework in detail, enabling the combination of multiple analyses, supporting reinterpretation and reuse (both within and outside collaborations), facilitating education on statistical methods, and advancing tool development by offering real-world examples for testing and debugging. This note summarizes the recent developments in the CMS experiment regarding the publication of statistical models.

2. Publishing with the CMS Combine tool

In December 2023, the CMS Collaboration decided to publish statistical models for all forthcoming analyses by default, in alignment with the open access policies of CERN and CMS, ensuring they are well-documented and comprehensible. While publishing statistical models has long been a desirable goal, it was previously hindered by the absence of a sufficiently robust technical infrastructure. This challenge was overcome with the public release of the CMS Combine tool.

Combine [3] is the statistical analysis software used within CMS, built on the ROOT, RooFit, and RooStats packages. It offers a command-line interface for common workflows in HEP statistical analysis, particularly those recommended by the CMS Statistics Committee. At the core of Combine is the *datacard*, a human-readable configuration file encapsulating the likelihood. Combine supports both predefined models and custom model-building. It is a powerful tool for combinations, scales well with model complexity, and offers extensive validation tools.

Combine can be compiled either within a CMS software (CMSSW) environment, which provides a versioned set of dependencies, or as a standalone package. Additionally, a precompiled version is available as a Docker container for non-CMS users, which can be installed as:

```
docker run [--platform linux/amd64] --name combine -it
gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1
```

Likelihoods for various analyses are published as COMBINE datacards, adhering to a set of nuisance parameter naming conventions adopted by the collaboration. The online documentation [4] provides up-to-date details on installation methods, analysis setup, running, statistical model and likelihood definitions, performing fits, statistical tests and diagnostic tools, and tutorials on main features.

3. Statistical models in COMBINE

The main task of COMBINE is to produce a statistical model, $p(\text{data}, \vec{\Phi})$, where $\vec{\Phi}$ are the model parameters. For numerical efficiency, this model can be expressed as

$$p(\vec{x}, \vec{y}; \vec{\Phi}) = p(\vec{x}; \vec{\mu}, \vec{\nu}) \prod_k p_k(\vec{y}_k; \vec{\nu}_k), \tag{1}$$

where it is factorized into a primary component constraining the POI $\vec{\mu}$ by the primary observables $\vec{x}$; and a secondary component constraining the nuisance parameters $\vec{\nu}$ by the auxiliary observable $\vec{y}$. The likelihood function can be constructed by evaluating $p(\vec{x}, \vec{y}, \vec{\Phi})$ on a data set:

$$\mathcal{L}(\vec{\Phi}) = \prod_d p(\vec{x}_d; \vec{\mu}, \vec{\nu}) \prod_k p_k(\vec{y}_k; \vec{\nu}_k) \tag{2}$$

where the subscript d runs over all entries in the data set. COMBINE implements a RooFit-based custom class to build the statistical model, which can subsequently be used in both frequentist and Bayesian calculations.

COMBINE hosts three default frequently used statistical models for the expression of the primary component $p(\vec{x}; \vec{\mu}, \vec{\nu})$. The first is a *counting analysis*, where only a single primary observable –the total event count n in a single channel– exists. The model is described by a Poisson probability

$$p(n; \lambda(\vec{\mu}, \vec{\nu})) = \lambda^n \frac{e^{-\lambda}}{n!}. \tag{3}$$

where λ represents the total number of expected signal and background events. Figure 1 shows an example COMBINE datacard for a counting experiment in a single channel. The data counts, signal and background processes with their expected yields, nuisance parameters representing systematic uncertainties affecting those processes with their constraints are all encoded in the datacard.

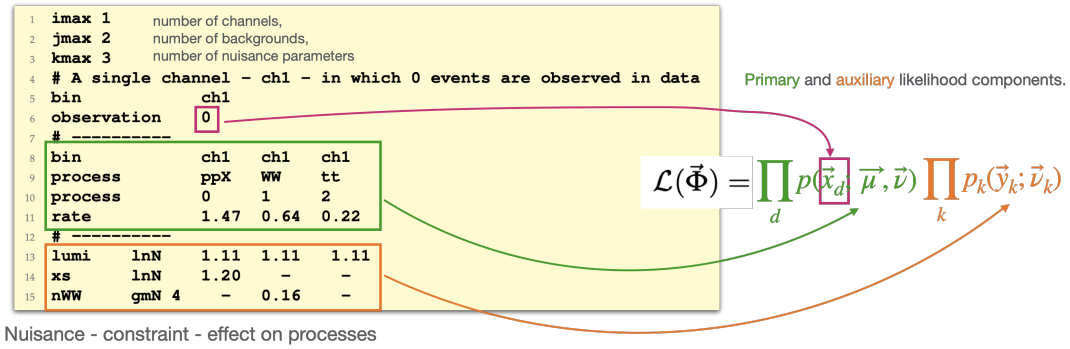

Figure 1: Example Combine datacard for a counting experiment in a single channel.

The second case is the *template shape analysis*, where the observable in each channel is partitioned into N_B bins. The statistical model in this case becomes a product of Poission probabilities for each of the included bins:

$$p(x;\vec{\mu},\vec{\nu}) = \prod_{b=1}^{N_B} P(n_b;\lambda(\vec{\mu},\vec{\nu})). \tag{4}$$

Inputs to this model, including data, nominal expectations and uncertainties as variations on expectations, are provided in histograms. Figure 2 left panel shows an example template shape analysis with data, signal and background processes, along with systematic variations on these processes. The template shape analysis is the most frequently used model in CMS.

The third case is the *parametric shape analysis*, where the model uses analytic functions rather than histograms to describe processes. Higgs boson measurements via a fit to the Higgs to diphoton mass distribution is a typical example of this case. Here, data can be input in a binned or unbinned manner, and the uncertainties on the expected distributions are uncertainties on the analytic function parameters. The parametric shape model can be expressed as

$$p(x;\vec{\mu},\vec{\nu}) = \sum_p \frac{\lambda_p(\vec{\mu},\vec{\nu}) f_p(x;\vec{\mu},\vec{\nu})}{\sum_p \lambda_p(\vec{\mu},\vec{\nu})} \tag{5}$$

where $f_p(x;\vec{\mu},\vec{\nu})$ are the probability density functions for each process p. Figure 2 right panel shows an example parametric shape analysis with data, functional distributions of signal and background processes, along with systematic variations on these based on variations on the function parameters.

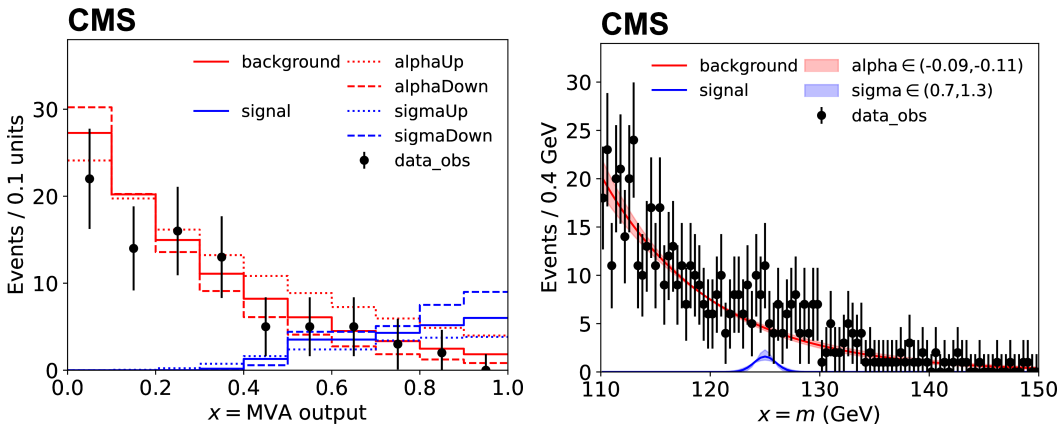

Figure 2: Left: An example template shape analysis with data, signal and background processes, along with systematic variations sigma and alpha, on these processes, respectively. Right: An example parametric shape fit on an invariant distribution. Data and functional distributions of signal and background processes are shown, along with systematic variations sigma and alpha on these processes based on variations on the function parameters.

Dedicated datacard formats also exist for the template shape and parametric shape models. In addition to these default models, Combine allows for the construction of custom statistical models, such as those involving multiple signal processes. This can be achieved by defining the POIs and their effects on the signal processes in a Python file.

4. Inference and diagnostic tools

The COMBINE package can perform a number of different statistical routines using the constructed statistical model by running the command line executable `combine` with different options:

```
combine <datacard.[txt|root]> -M <method>
```

where `method` specifies the statistical calculation to be performed. Below are the key COMBINE methods used for statistical inference:

- `HybridNew`: Computes modified frequentist limits with pseudo-data, p-values, significance, and confidence intervals with several options. The `--LHCmode LHC-limits` option is the recommended one.
- `AsymptoticLimits`: Calculates limits according to the asymptotic formulas in [5], valid for large event counts.
- `Significance`: simple profile likelihood approximation for calculating significances. `BayesianSimple` and `MarkovChainMC`: Compute Bayesian upper limits and credible intervals for simple and arbitrary models.
- `MultiDimFit`: Performs maximum likelihood fit with multiple POIs; estimates confidence intervals from likelihood scans.

COMBINE can perform goodness-of-fit tests, and is equipped with the following diagnostic tools:

- `GoodnessOfFit`: Performs goodness-of-fit (GoF) test for models including shape information using various estimators (i.e., saturated, Kolmogorov-Smirnov, Anderson-Darling)
- `Impacts`: Evaluate the shift in POI from $\pm\sigma_{postfit}$ variation for each nuisance parameter.
- `ChannelCompatibiltyCheck`: Checks how consistent the individual channels of a combination are.
- `GenerateOnly`: Generates random or Asimov pseudo-datasets for use as input to other methods.

5. The first public statistical models

The first statistical model published by CMS [6] is that belonging to the observation of the Higgs boson at a mass of 125 GeV with 7 and 8 TeV data collected during 2011 and 2012. The search was performed in five decay modes: gamma gamma, ZZ, WW, tau tau, and b b-bar. The model is published in the CERN Document Server, and is accompanied by detailed instructions for combining the five channels, calculating the significance, measuring the signal strength, and building a new model that has Higgs boson-vector boson and Higgs boson-fermion coupling modifiers as POIs. More recently, a second statistical model was published for the measurement of inclusive and differential cross sections for W^+W^- production with 13.6 TeV with data collected in 2022 [7].

More models, especially those involving searches for physics beyond the standard model (BSM), will be published to support reinterpretation efforts. BSM models pose challenges due to their multiple parameters and extensive scans, leading to hundreds or thousands of signal points per analysis, each with specific values for uncertainties. A compact way of publishing these is to provide

a single template datacard accompanied by interpolated functions for signal rates and uncertainties in terms of signal model parameters. Interpolation would be performed using `RooSplineND` for counting-style analyses or generated automatically through keyword input for shape analyses. An open question remains regarding the treatment of uncertainties when dealing with entirely new signal models.

6. Conclusions and outlook

CMS has released its first statistical models implemented in Combine and is in the process of publishing more. To streamline future releases, upcoming CMS analyses are adopting a standardized nuisance parameter naming convention. The Combine tool is now publicly available, along with comprehensive documentation and a standalone container, providing users with a self-documenting statistical model-building framework and an extensive toolset for statistical inference. Efforts are also underway to enhance interoperability with other formats. A Combine-to-pyHF conversion tool has been developed and thoroughly validated, and work has begun on implementing the HEP Statistics Serialization Standard (HS3) for broader compatibility.

Acknowledgements: SS is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under contracts NRF-2021R1I1A3048138, NRF-2018R1A6A1A06024970, and NRF-2008-00460.

References

[1] F. James, Y. Perrin and L. Lyons, "Workshop on confidence limits, CERN, Geneva, Switzerland, 17-18 Jan 2000: Proceedings," doi:10.5170/CERN-2000-005

[2] K. Cranmer, S. Kraml, H. B. Prosper, P. Bechtle, F. U. Bernlochner, I. M. Bloch, E. Canonero, M. Chrzaszcz, A. Coccaro and J. Conrad, *et al.* "Publishing statistical models: Getting the most out of particle physics experiments," SciPost Phys. **12** (2022) no.1, 037 doi:10.21468/SciPostPhys.12.1.037 [arXiv:2109.04981 [hep-ph]].

[3] A. Hayrapetyan *et al.* [CMS Collaboration], "The CMS Statistical Analysis and Combination Tool: COMBINE", [arXiv:2404.06614 [physics.data-an]].

[4] "CMS Combine web documentation", https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/

[5] G. Cowan, K. Cranmer, E. Gross and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," Eur. Phys. J. C **71** (2011), 1554 [erratum: Eur. Phys. J. C **73** (2013), 2501] doi:10.1140/epjc/s10052-011-1554-0 [arXiv:1007.1727 [physics.data-an]].

[6] CMS Collaboration, "CMS Higgs boson observation statistical model (v1.0)", CERN. https://doi.org/10.17181/c2948-e8875

[7] CMS Collaboration, "Statistical models for CMS-SMP-24-001", CERN. https://doi.org/10.17181/bp9fx-6qs64