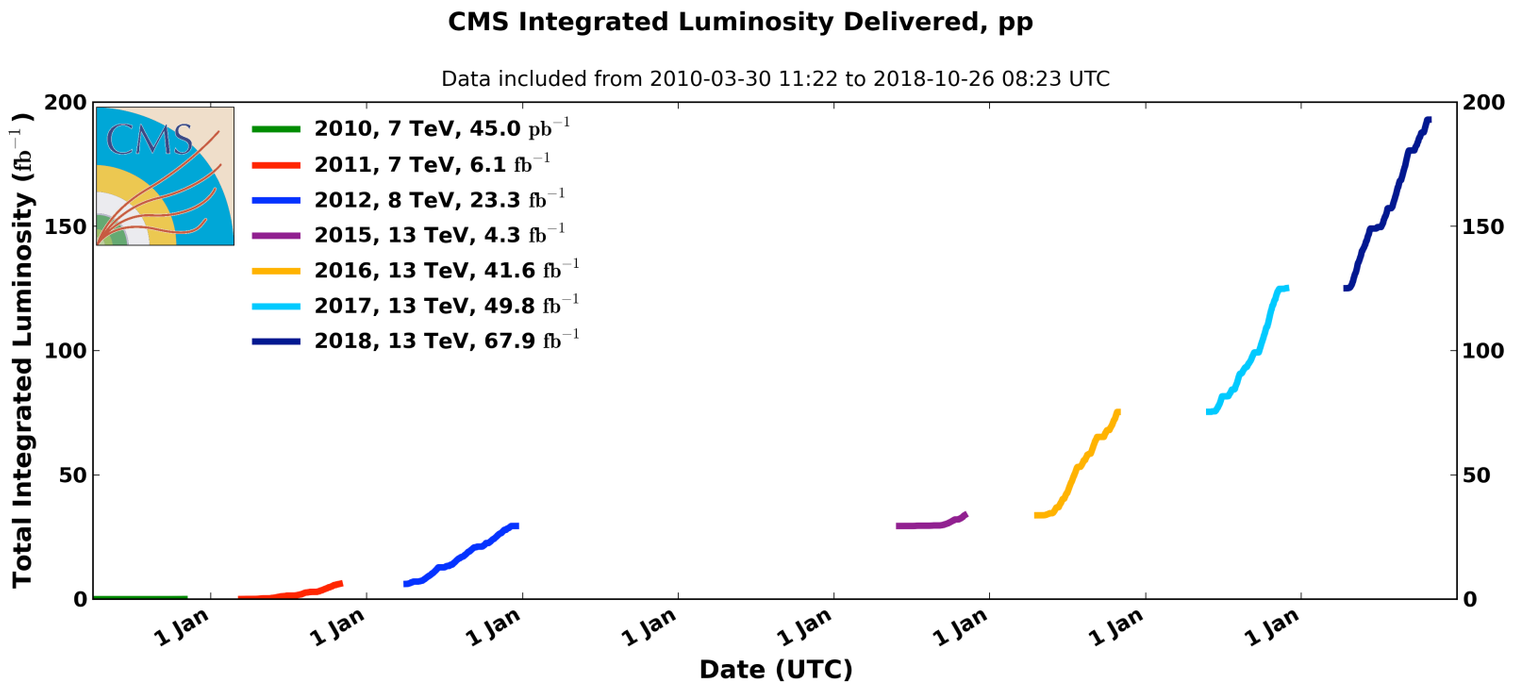
Review of statistical practices in CMS

Andrea Carlo Marini

Virtual Tribute to QCHS 2021 — Statistics session

Introduction

- Statistical practices consolidate through out the years
- 10y of data-taking



- More and more awareness of limitations
- Increasing cpu powers and capabilities
- Standardisation of tools *HiggsCombine*

- Not entering in the details of any analysis

The Higgs discovery — practices

- Higgs search and discovery was a plain field for assessing and agreeing on common methodologies between the ATLAS and CMS Collaborations

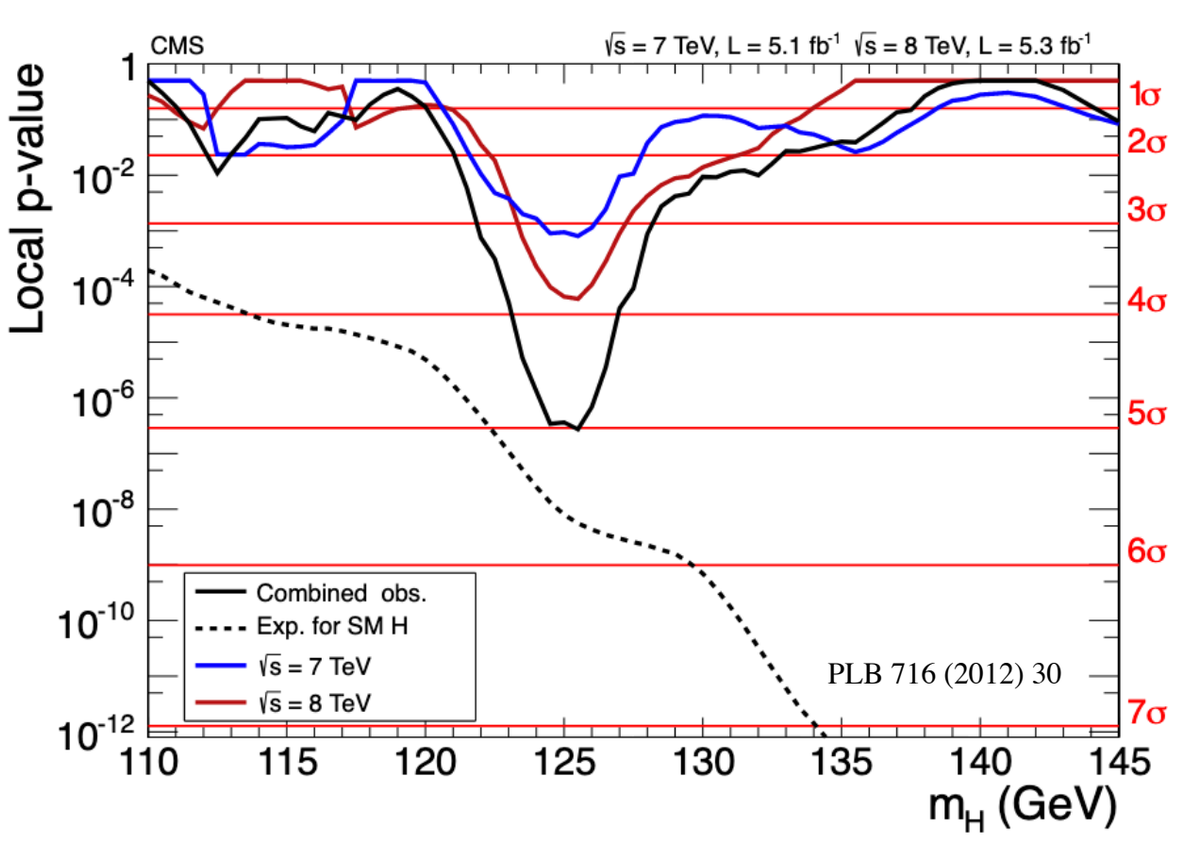
ATL-PHYS-PUB-2011-11

CMS NOTE-2011/005

- The methodology of the search and the presentation of the results was agreed upon in 2011

- Cumulative improvements over time reflected the availability of common tools and larger computing power

- Frequentist methods usually preferred over Bayesian inference



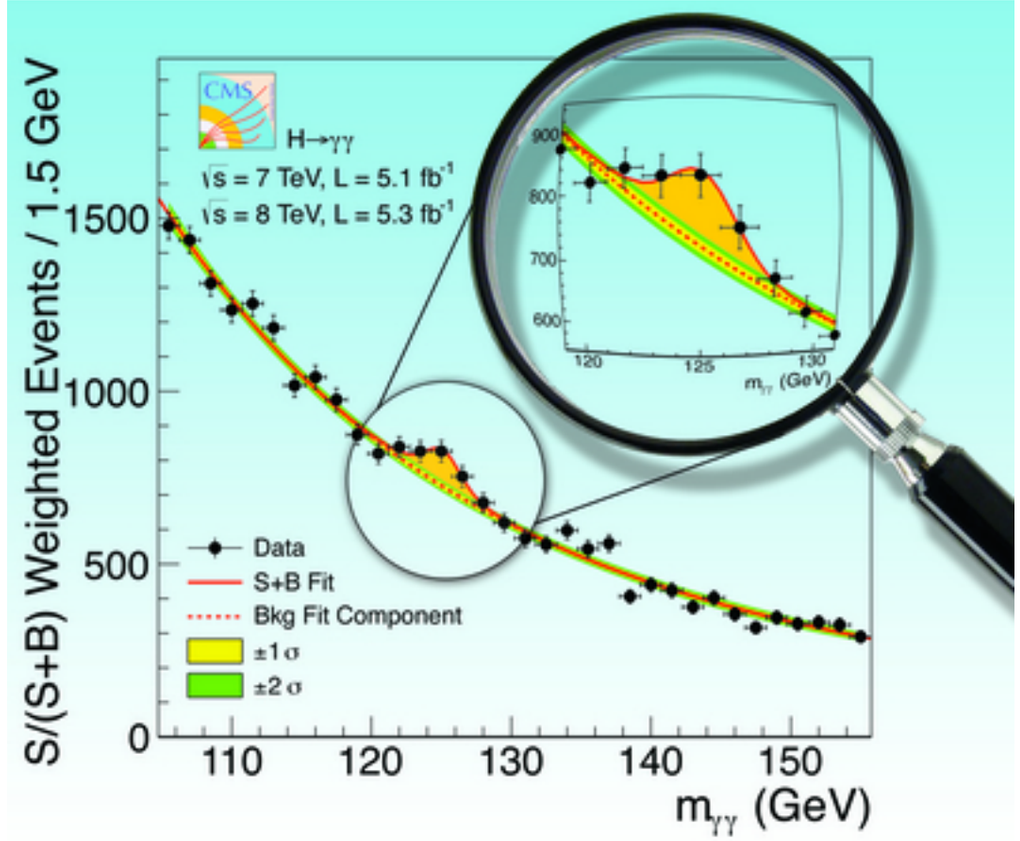
Blind analysis

- Analysis selection, optimisation, training of discriminators, performances … performed without looking at the actual observation

- Analysis strategy frozen, results are derived in the Signal Regions

- Control over LEE deriving from requirement choices

- Clean interpretation in terms of p-value



The limit setting & discovery

- Test statistics: Profiled likelihood ratio
  - Nuisances are profiled for different values of the parameters of interest

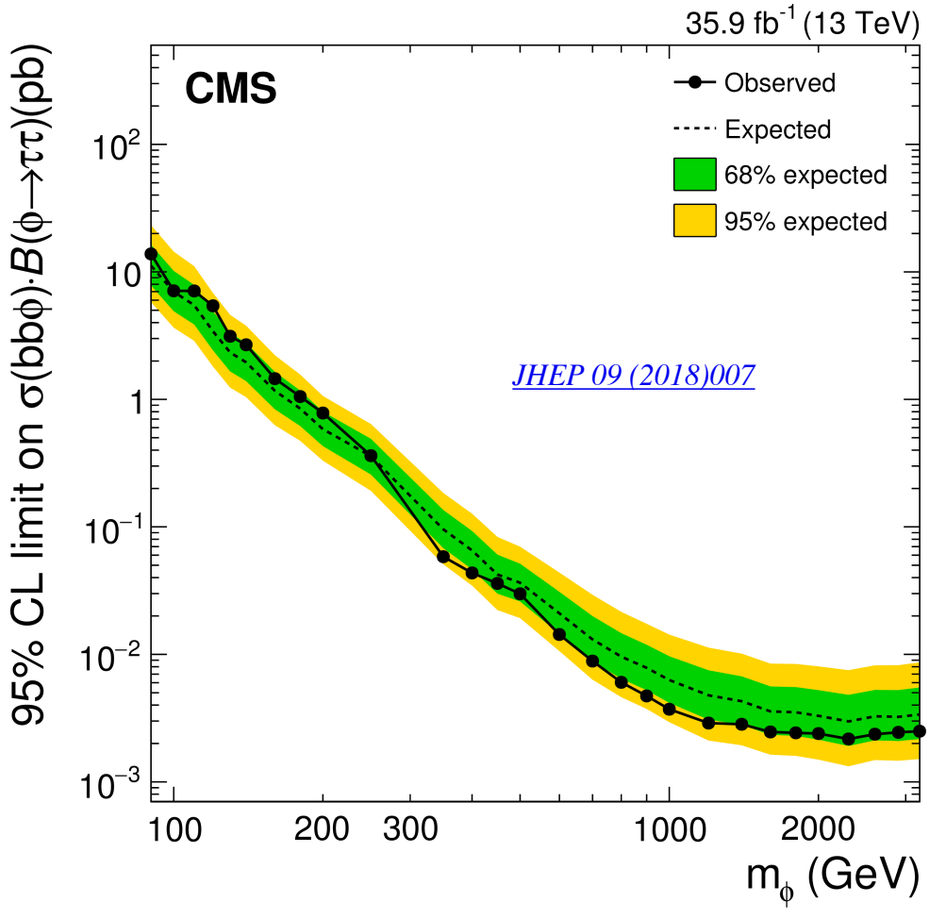
- Modified test statistics for upper limit

$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}, \quad \text{with a constraint } 0 \leq \hat{\mu} \leq \mu$$

- CLs criterion (95%CL) *J.Phys.G 28 (2002) 2693*, *NIMA 434 (1999) 435*

$$CL_s(\mu) = \frac{p_\mu}{1 - p_b}$$

- Asymptotic formulae *EPJC 71 (2011) 1554*



Quantifying excesses

- Excesses are quantified as a p-value based on the likelihood ratio test statistics between the signal+background and background only hypothesis

$$q_0 = -2 \ln \frac{\mathcal{L}(\text{data}|0, \hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})} \quad \text{and } \hat{\mu} \geq 0.$$

- P-values are reported in Z-score units using the one-sided convention

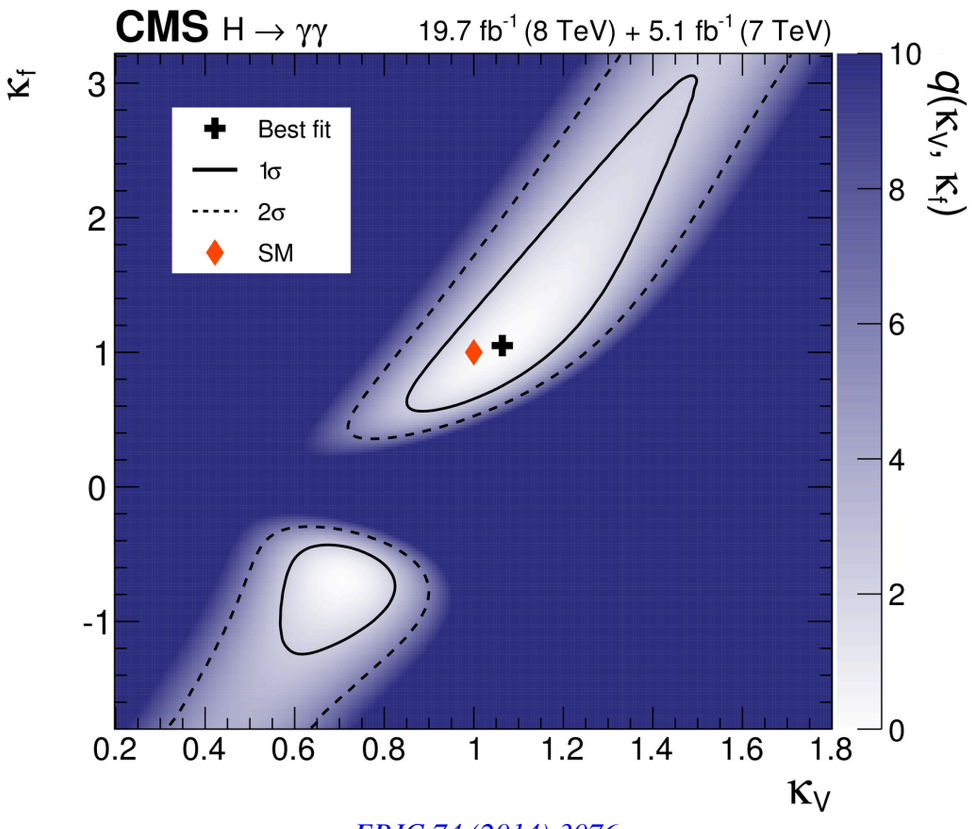
$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx = \frac{1}{2} P_{\chi^2_1}(Z^2),$$

- Asymptotic behaviour based on Wilks' theorem is often assumed [Ann. Math. Statist. 9 (1938) 60]

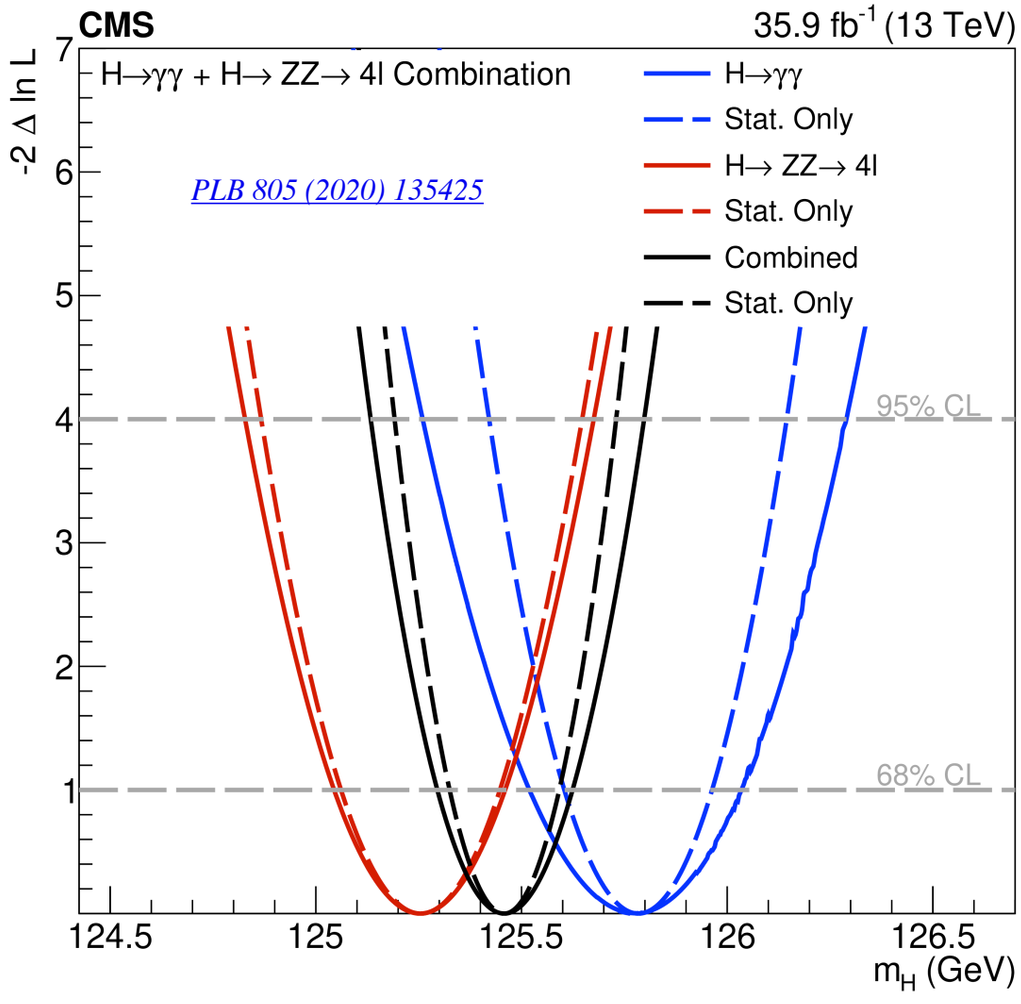
$$p^{estimate} = \frac{1}{2} \left[1 - \text{erf} \left(\sqrt{q_0^{\text{obs}}/2} \right) \right].$$

Measurement of parameters

- Log-likelihood ratio and Wilks' theorem is the most used.
- Derivation of Confidence intervals with pseudo experiments using Feldman-Cousins method with Cousins-Highland prescription for systematic treatment.



EPJC 74 (2014) 3076

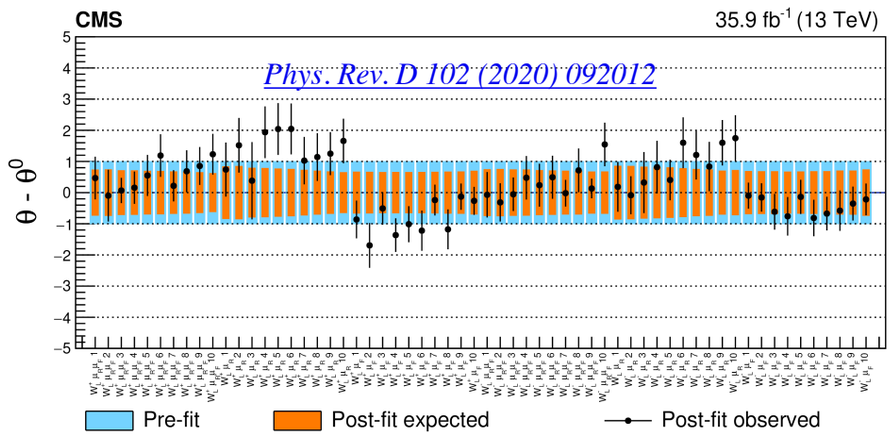


Uncertainties & nuisance parameters

- Uncertainties are often incorporated in results as nuisance parameters
  - pdf: Gaussian, LogNormal, Gamma …
- Uncertainties derived from MC are interpolated and a gaussian pdf is assumed.
- Relevant nuisances are correlated / uncorrelated in the different final states

- Either profiled or used with the Hybrid bayesian-frequentist method
- Constrained nuisances are studied in order to establish the genuinely of the constrain term
  - Spurious constrain may araise from reduced statistical mc precision

- Uncertainties due to finite Monte Carlo statistics usually added with the Barlow—Beeston lite approach
  - For non-likelihood methods, pseudo-experiments are often used to propagate this uncertainty



Non-Asymptotic results

- Based on pseudo-experiments
- Possible with ever-increasing computing power
- Usually performed either as main results or as a cross check of the asymptotic formulae, in points with small numbers of events in the dataset.
  - Depending on the settings, this can make a sizeable effect in the results
- For the toy generation, the nuisance parameters are fixed to their post-fit values from the data, while the constraint terms are randomised in the evaluation of the likelihood.
  - Limit settings (as the asymptotic ones) use the modified test statistics and constraining POIs to positive values

Feldman-Cousins confidence intervals

Look Elsewhere Effect

- Many analyses concern a very wide parameter space defined by a group of models
- Analysis presents both local and global p-values, correcting for the different possible models in the analysis

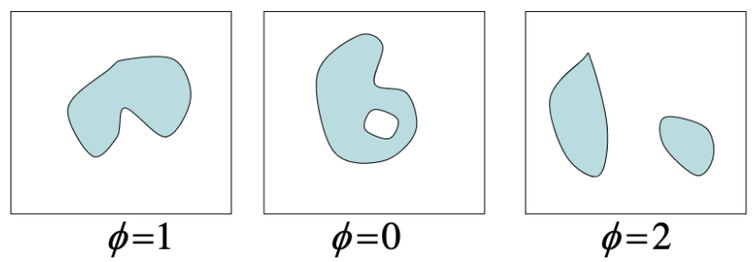
$$N = \frac{\log(1 - p_{\text{global}})}{\log(1 - p_{\text{loc}})} \approx \frac{p_{\text{global}}}{p_{\text{loc}}}$$

- Moderate values of global p-values are derived via pseudo-experiments
- Extrapolation to very large significance levels 1D (2D) using likelihood crossings (Euler characteristics)

$$\langle N_u \rangle \;=\; \langle N_{u_o} \rangle \, e^{-(u-u_o)/2},$$

EPJC 70 (2010) 525

Astr. Phys. 35 (2011) 230



Quantifying model agreement

To condense in one number the agreement of predictions with data, often choose

- Chi2 test
- Kolmogorov-Smirnov
- Generalized chi2 (saturated likelihood), *Note*

In likelihood-based methods the distribution of the test statistics is derived from pseudo experiments

Unfolding of spectra

Unfolding is used to “unsmear” effects due to detector resolution and efficiencies. Based on chi2 or full likelihood (simultaneously with signal extraction)

$$\vec{x}_{\text{reco}} = \hat{\mathbf{R}} \cdot \vec{x}_{\text{true}} + \vec{b}$$

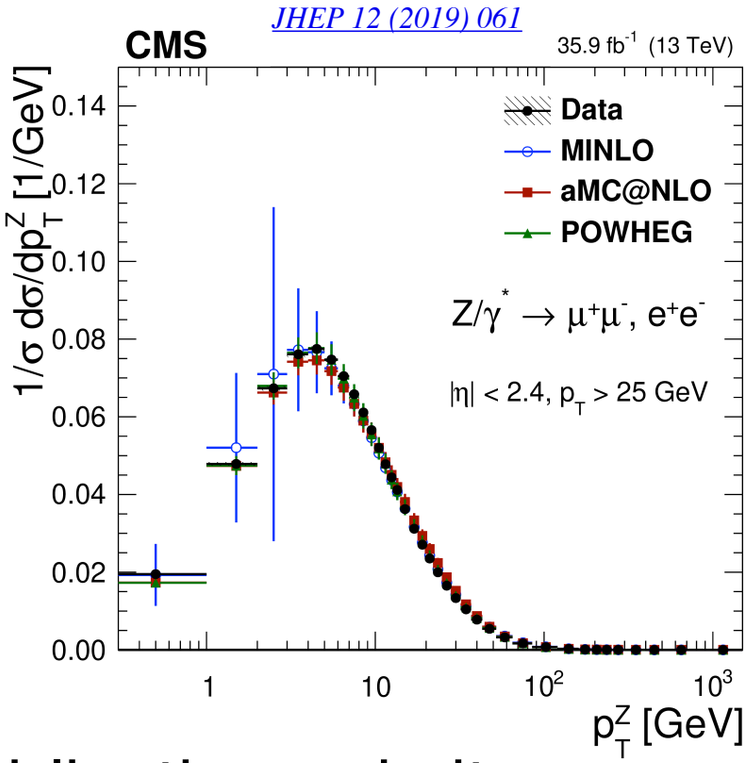
$$\mathcal{L} = \prod_{i \in \text{reco}} \mathcal{P}(x_{\text{reco},i} | \sum_{j \in \text{gen}} \mu_j \mathbf{R}_{ij} + \vec{b}_i)$$

Regularisation types used the most:

- None
- d’Agostini
- Tikhonov (SVD, or TUnfold)

Most optimised regularisation methods:

- L-Curve
- Minimum Global correlation coefficient



Covariance matrices usually included in the publication or in its digitalisation (hepdata)

Summary

- The CMS Collaboration has been refining, and is still improving to this day, ways of analysing data and presenting the results.

- There is very significant standardisation of the statistical methods used for the analysis and presentation of results, as well as for the crosschecks that assess the validity of any assumptions used.

- Throughout the years there is a tendency to use more the likelihood function, including shape analysis and control regions

- Nowadays, more attention to ensuring the preservation of the published results and the potential for their reinterpretation; covariance matrices and Rivet routines are increasingly being used in publications.