Statistical methods at ATLAS and CMS

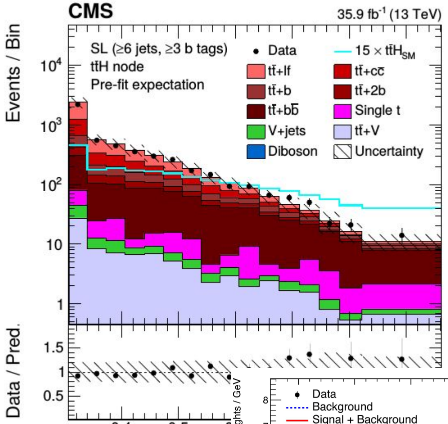
Michele Pinamonti (ATLAS)
University and INFN Roma "Tor Vergata"

Higgs Toppings Workshop, Benasque 28th May - 1st June 2018

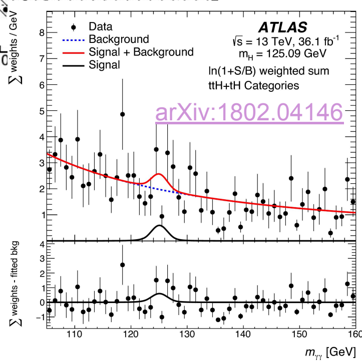
Building Likelihoods

- Most analyses (especially ttH/tH, with more data available...) are "***shape analyses***"
  - based on **distributions** of continuous observables
  - signal and background predictions depend on **parameters**:
    - parameter of interest, **POI** (e.g. signal strength μ)
    - eventually other ("nuisance") parameters, **NPs** *(e.g. background normalization...)*

arXiv:1804.03682



- Build a ***global likelihood function***:
  - binned likelihood: $$\mathcal{L}(\boldsymbol{n}|\mu,\boldsymbol{\theta}) = \prod_{i\in bins} \mathcal{P}(n_i|\mu\cdot S_i(\boldsymbol{\theta}) + B_i(\boldsymbol{\theta}))$$
    Poisson; observed bin contents; parameters; *S*+*B* prediction in bin *i*
  - unbinned likelihood: $$\mathcal{L}(\boldsymbol{m}|\mu,\boldsymbol{\theta}) = \mathcal{P}(n_{obs}|S+B) \times \prod_{i=0}^{n_{obs}} \frac{S\cdot P_S(m_i,\boldsymbol{\theta}) + B\cdot P_B(m_i,\boldsymbol{\theta})}{S+B}$$
    values of observable *m*; PDFs for for *S* and *B*



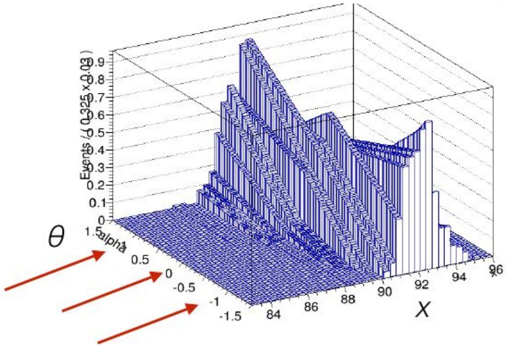
arXiv:1802.04146

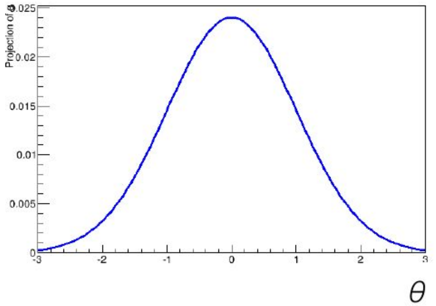
- Result = POI value that maximizes the likelihood

The Profile Likelihood approach

- The profile likelihood is a way to include **systematic uncertainties in the likelihood**
  - systematics included as "***constrained***" **nuisance parameters**
  - the idea behind is that systematic uncertainties on the measurement of μ come from ***imperfect knowledge*** of parameters of the model (*S* and *B* prediction)
    - still *some knowledge* is implied: "$\theta = \theta_0 \pm \Delta\theta$"

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\theta}^0 | \mu, \boldsymbol{\theta}) = \prod_{i \in bins} \mathcal{P}(n_i | \mu \cdot S_i(\boldsymbol{\theta}) + B_i(\boldsymbol{\theta})) \times \prod_{j \in syst} \mathcal{G}(\theta_j^0 | \theta_j, \Delta\theta_j)$$





- usually θ^0=0 and $\Delta\theta$=1 (convention)
- define **effect of systematic** *j* on prediction *x* in bin *i* at "+1" and "-1",
- then interpolate & extrapolate for any value of θ

  - external / *a priori* knowledge interpreted as "**auxiliary/subsidiary measurement**", implemented as **constraint/penalty term**, i.e. probability density function (*usually Gaussian, interpreting "±Δθ" as Gaussian standard deviation*)

Normalization factors and MC statistics

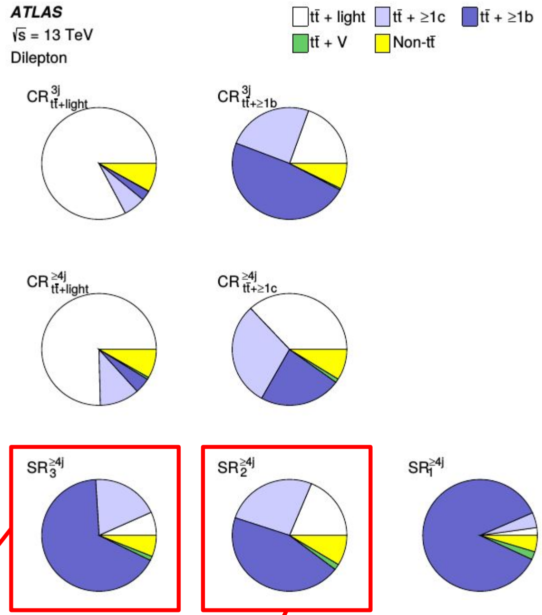
- Beside NP associated to systematic uncertainties,
  other NP can be included in the likelihood as ***free parameters***, in the same way as the POI:
    - called "**normalization factors**" (NF)
    - **no *prior***, *multiplicative* factors (⇒ linear) for particular ***S***, ***B*** or ***B*** components:

$$B(\theta, k) = k \cdot B(\theta)$$

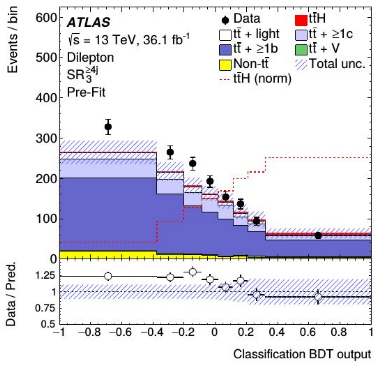
- Statistical uncertainty from **limited number of** (MC) **events** used to build the histograms for predicted ***S*** and ***B*** result in **independent uncertainties in each bin**, referred to as "**MC stat.**"
    - implemented as additional NPs (one per bin) with scaled Poisson ("*gamma*") priors
    - default: single MC-stat NP assigned to total prediction (***S***+***B***) in each bin:
        - problematic for signal, or in general component of the prediction with the POI attached to it (e.g. if these NP get pulled)
          ⇒ NOT applied to signal
        - could consider to **split it**: useful?

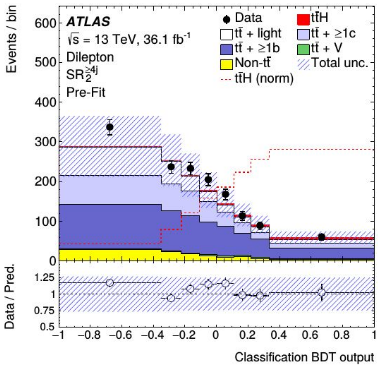
Categories, SRs, CRs...

- Multiple **analysis regions** / **event categories** often used:
  - decay modes
  - kinematic selections
- Useful to model these **separately** if
  - **sensitivity** is better in some regions (*avoids dilution*)
  - some regions can ***constrain*** NPs (*including NP for systematics*)
    - e.g. ***control regions*** for backgrounds
- Analyse them simultaneously to model **correlations** between the regions (*common NPs*)
  - better than a-posteriori combination



Phys. Rev. D 97, 072016





- If **no statistical correlations** (*orthogonal*)
  ⇒ can simply take **likelihood product**
- Usually **coherent/correlated effect** of all parameters (*including NP for systematics*) in all bins, in all regions

Profile Likelihood maximization

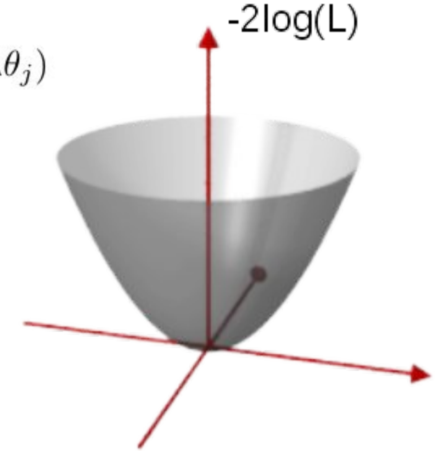
- With such a likelihood defined, the measurement of the parameter of interest (POI, or *μ*) becomes a ***N-dimensional likelihood maximisation*** (*or negative-log-likelihood minimization*) problem:

N = N(POI) + N(NP)

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\theta}^0 | \overbrace{\mu, \boldsymbol{\theta}}) = \prod_{i \in bins} \mathcal{P}(n_i | \mu \cdot S_i(\boldsymbol{\theta}) + B_i(\boldsymbol{\theta})) \times \prod_{j \in syst} \mathcal{G}(\theta_j^0 | \theta_j, \Delta\theta_j)$$

  - ⇒ N-dimensional fit
    - fit result is "best point" *(μ,* ***θ****)*

$$(\hat{\mu}, \hat{\theta}_0, ...\hat{\theta}_{N-1}) : \mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}) = max$$



“best-fit” point (μ,θ)

Profile Likelihood Ratio

- Neyman-Pearson lemma:
  - the **likelihood ratio** $L(H_0)/L(H_1)$ is the **optimal discriminator** when testing hypothesis H_1 vs. H_0 - e.g. H_1 = presence of signal (μ>0), H_0 no signal (μ=0)
- In case of profile likelihood, define ***profile likelihood ratio*** (PLR):

Profile likelihood ratio only dependent on μ

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\theta}}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\theta})}$$

Maximize L for a given μ 'conditional' likelihood

Maximize L 'unconditional' likelihood

- **Test statistics** defined as:

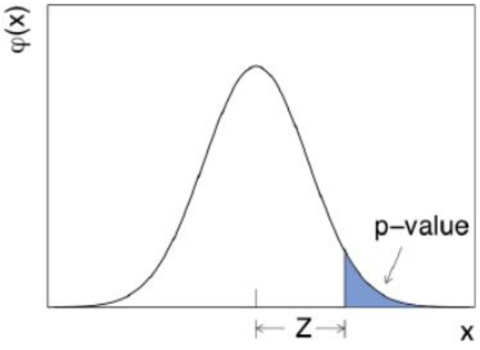
$$t_\mu = -2 \ln \lambda(\mu) \quad , \quad q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases} \quad , \ldots$$

- Can then build ***p-value*** and ***significance***:

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

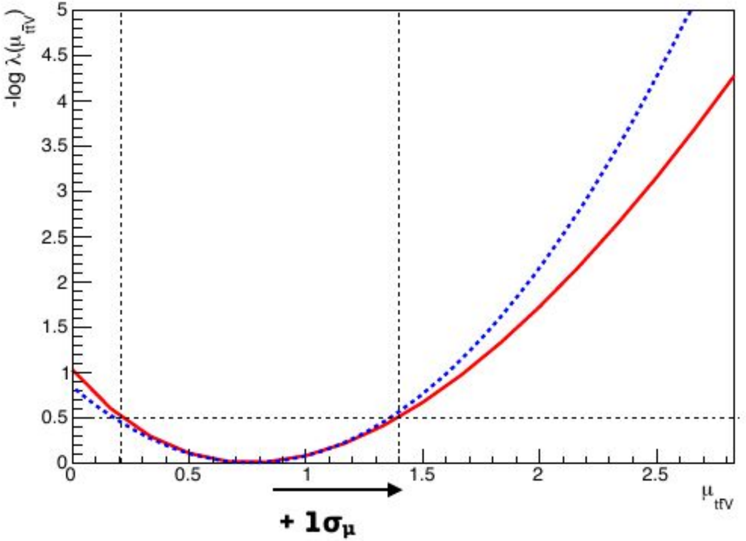
$$Z_0 = \Phi^{-1}(1 - p_0)$$

distribution of test statistics



Asymptotic regime

- In **large statistics** data samples, the *distribution* of the test statistic is **known** according to Wilks' Theorem (*independently on the prior!*):
  - **χ2 distribution**
  - parabolic shape around the minimum



$$-2\log\lambda(\mu) = -2(\log L(\mu,\hat{\hat{\theta}}) - \log L(\hat{\mu},\hat{\theta})) = \left(\frac{\mu-\hat{\mu}}{\sigma_\mu}\right)^2$$

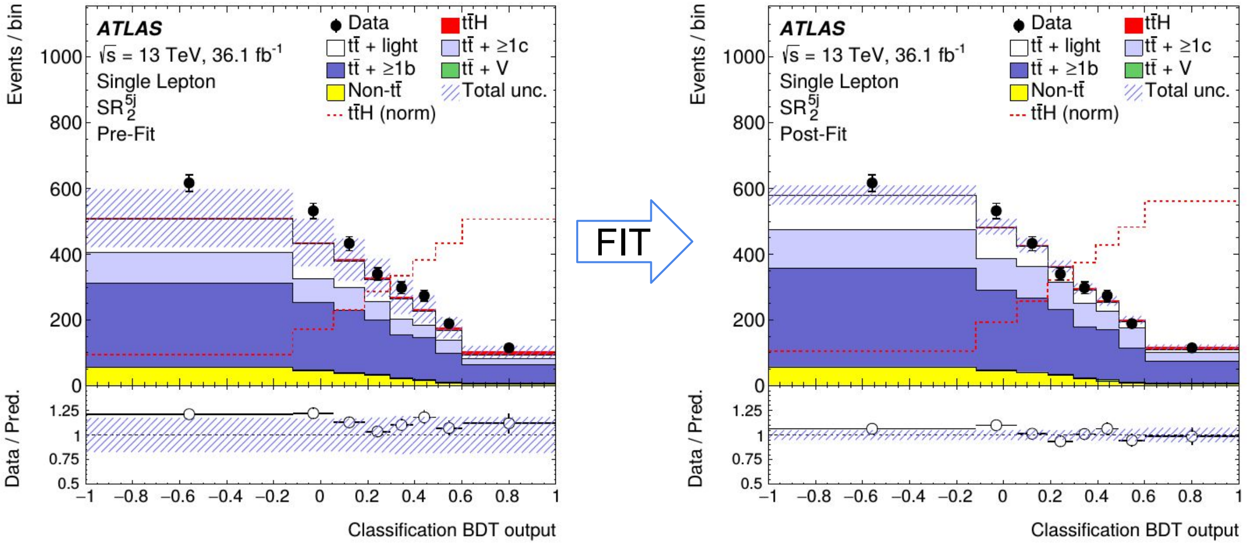
⇒ can directly calculate p_0 ⇒ **significance**

⇒ can get the **uncertainty on** μ

- No need to use pseudo experiments
- This theorem holds true for even as few as ~ O(10) events in a data sample

Profiling, pre-fit and post-fit

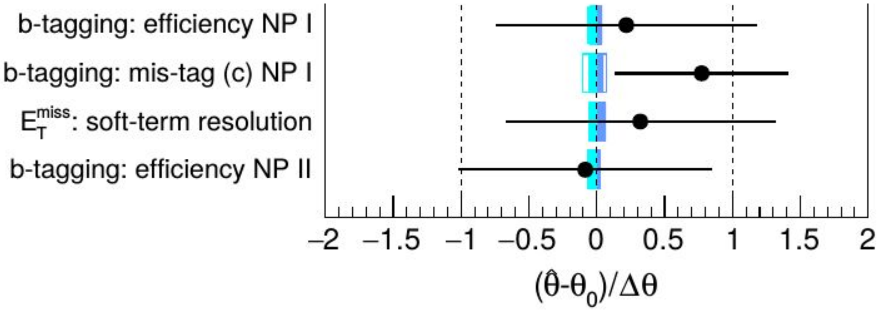
- Profile likelihood fit can:
  - **change background prediction**, if best-fit θ values different from θ_0
  - **reduce uncertainty** on background, through:
    - ***constraint*** of NPs
      (*"improved knowledge" of parameters that are affected by systematic uncertainties, i.e. data have enough statistical power to further constraint the NP*)
    - ***correlations*** between NPs

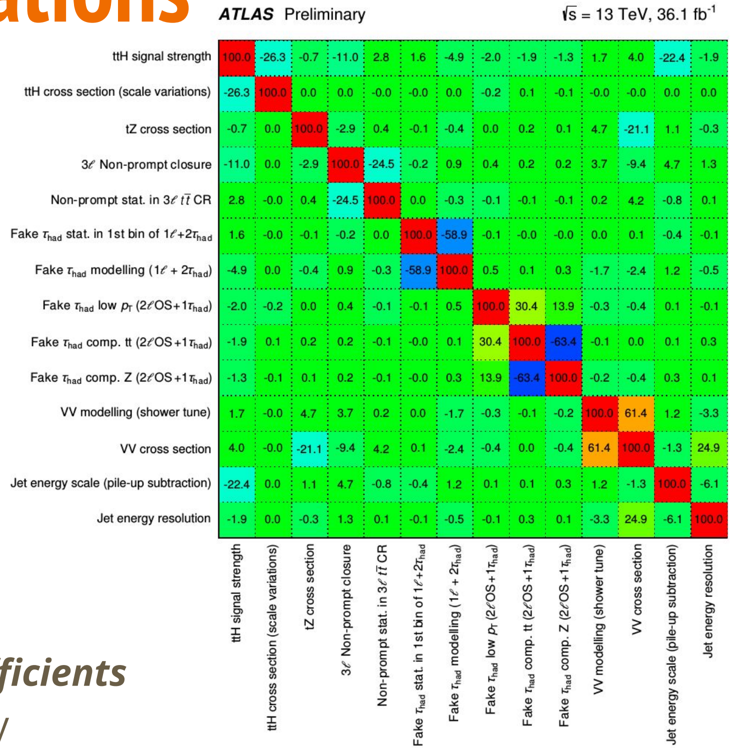


NP pulls, constraints and correlations

Phys. Rev. D 97, 072003

- Useful to **monitor** NP **pulls** and **constraints**:
  - they are "*nuisance*", but they are important!



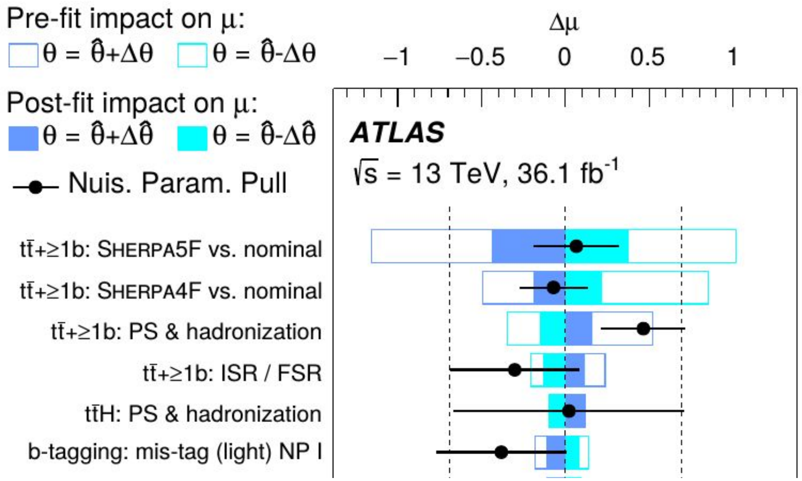


- Important to consider also NP **correlations**:
  - uncertainties on NPs (*and POI*) extracted from ***covariance matrix***, which includes ***correlation coefficients***
    - correlation **built by the fit**, even if completely independent / uncorrelated sources of uncertainty before the fit (*correlation in the improved knowledge of the parameters*)
    - (anti-)correlations can **reduce** total post-fit uncertainty!

Impact of NP on the POI

- To answer the question "***which systematics are more important?***"

- The "**ranking plot**" shows *pre-fit* and *post-fit* **impact** of **individual NP** on the determination of μ:
    - **each NP fixed** to ± 1 pre-fit and post-fit sigmas ($\Delta\theta$ and $\Delta\hat{\theta}$ = uncertainty on $\hat{\theta}$)
    - fit re-done with *N-1* parameters
    - impact extracted as difference in **central value** of μ



Splitting of uncertainties

- To answer a similar but different question:
  - **how much of the total uncertainty** comes from a certain ***set of*** systematic uncertainties?
  - or similarly, how large is the pure "statistical uncertainty"?

- Procedure:
  - **fix a group** of NPs to post-fit values
  - repeat the fit
  - look at **error on μ** this time and get $\Delta\mu$ as quadratic difference between full and reduced error
  - statistical uncertainty obtained by fixing all NPs

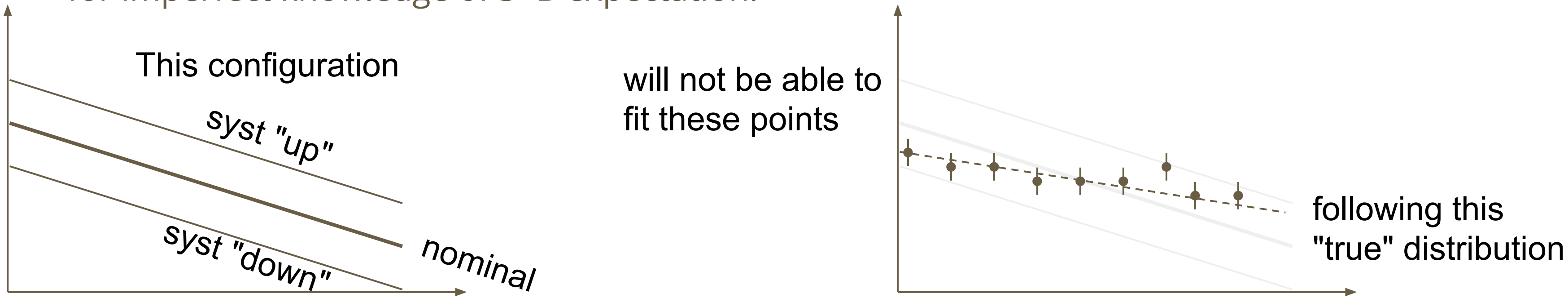
arXiv:1804.03682

Uncertainty source	$\pm\Delta\mu$ (observed)	$\pm\Delta\mu$ (expected)
Total experimental	$+0.15/-0.16$	$+0.19/-0.17$
b tagging	$+0.11/-0.14$	$+0.12/-0.11$
jet energy scale and resolution	$+0.06/-0.07$	$+0.13/-0.11$
Total theory	$+0.28/-0.29$	$+0.32/-0.29$
$t\bar{t}$+hf cross section and parton shower	$+0.24/-0.28$	$+0.28/-0.28$
Size of the simulated samples	$+0.14/-0.15$	$+0.16/-0.16$
Total systematic	$+0.38/-0.38$	$+0.45/-0.42$
Statistical	$+0.24/-0.24$	$+0.27/-0.27$
Total	$+0.45/-0.45$	$+0.53/-0.49$

Profiling issues

- The profile likelihood approach is **valid** with some **assumptions**
  - in particular, assumed that "*nature*" can be described by the model with ***a single combination of values*** for the parameters

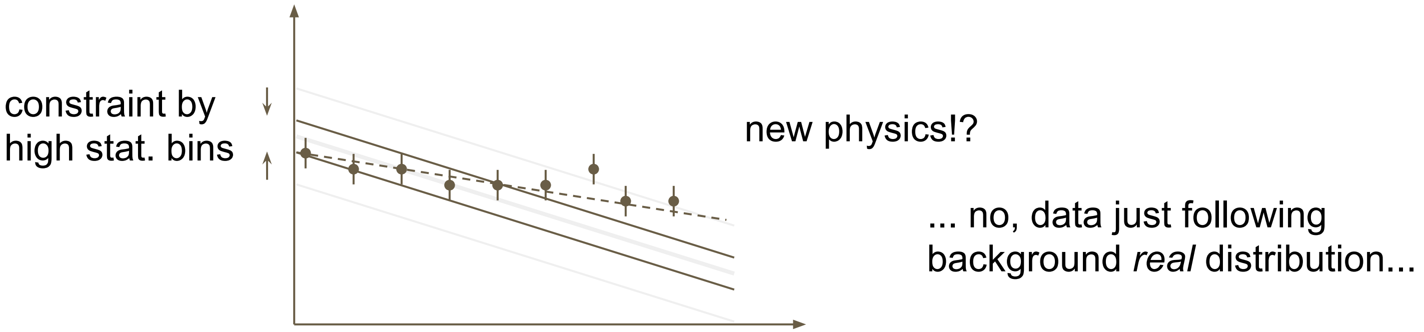
- Cannot just take *large uncertainties* hoping that they are enough to cover for imperfect knowledge of S+B expectation!



- "**Flexibility**" / "**granularity**" of the systematics model needs to be considered

The constraint issue

- Flexibility more and more **critical** when **statistical uncertainty** on data becomes less and less important w.r.t. systematics
  - e.g. taking the example before:



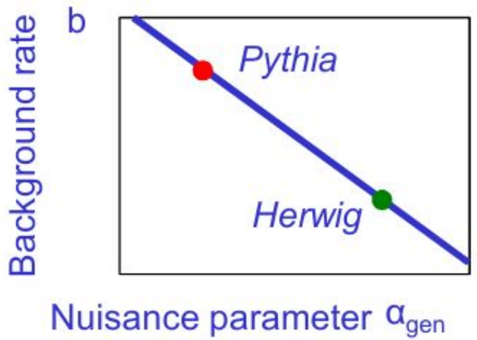
- More real examples:
  - **single JES** systematic NP across all jet energy spectrum allows high-stats low-energy control regions/bins to calibrate JES for high energy jets → intended?
  - **simple ± 50%** overall uncertainty on *tt*+jets background, probably enough to cover uncertainties also in remote phase-spaces (*tails of distributions for tt+HF-enriched selection*), but data in *tt*+light-jets-enriched CRs will constrain it to <5%, propagated to SRs… → ok?

Theory modeling systematics

- ***Experimental systematics*** nowadays often well suited for profile likelihood application:
  - come from calibrations ⇒ gaussian constraint appropriate
  - broken-down into several independent/uncorrelated components (JES, *b*-tagging...)
- Different situation for **theory systematics**:
  - **difficulty 1:** what is the **distribution** of the subsidiary measurement?
  - **difficulty 2:** what are the **parameters** of the systematic?
    - can a combination of the included parameters describe **any possible** configuration?
    - is **any allowed value** of the parameter physically meaningful?

See: https://indico.cern.ch/event/287744/contributions/1641261/attachments/535763/738679/Verkerke_Statistics_3.pdf

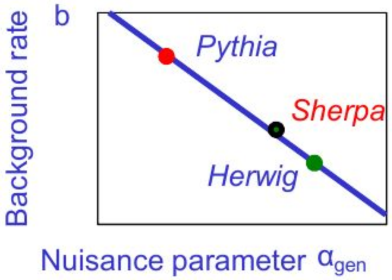
- ***The obviously tricky case:* "two point" systematics**
  - e.g. Herwig vs. Pythia as "parton shower and hadronization model uncertainty", as a single NP



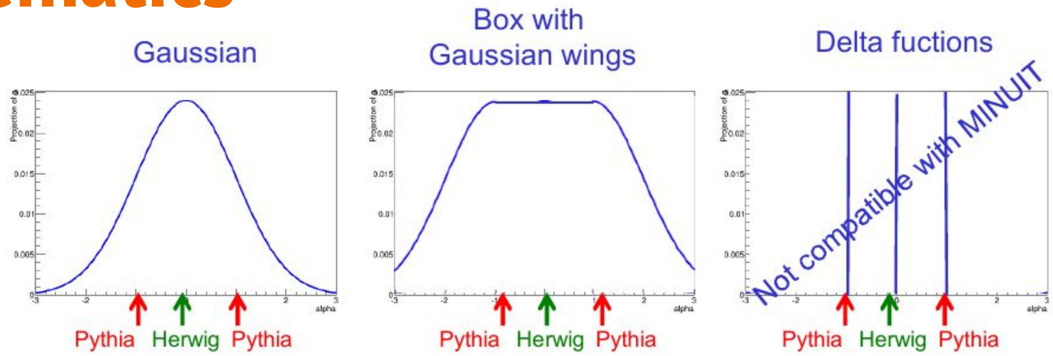
Theory modeling systematics

One-bin case:

- reasonable to think that "Sherpa" can be between Herwig and Pythia



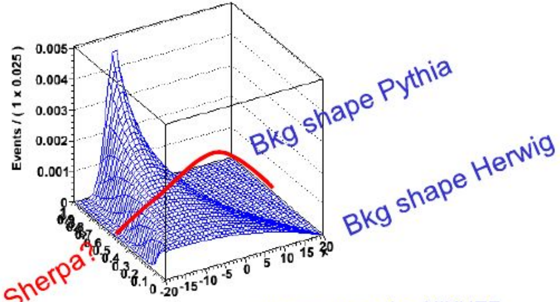
Which prior?



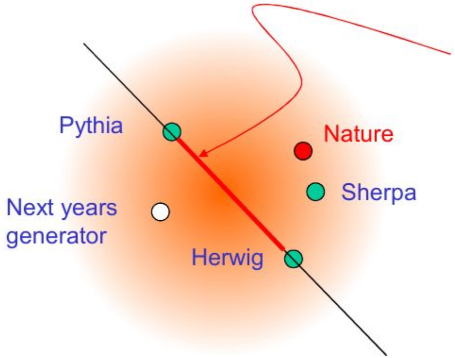
Pre-fit / non-constrained NP could be fine to cover for all possible models...

Shape case:

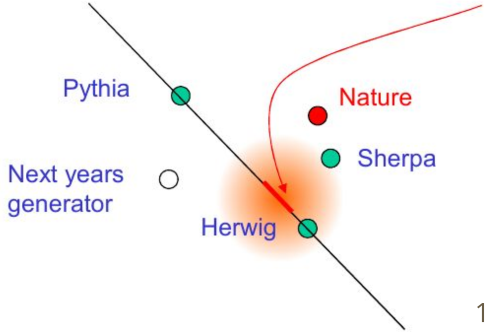
- Sherpa can be different from linear combination of Py and Her...



Wouter Verkerke, NIKHEF



... but is this level of constraint ok?



Theory modeling systematics

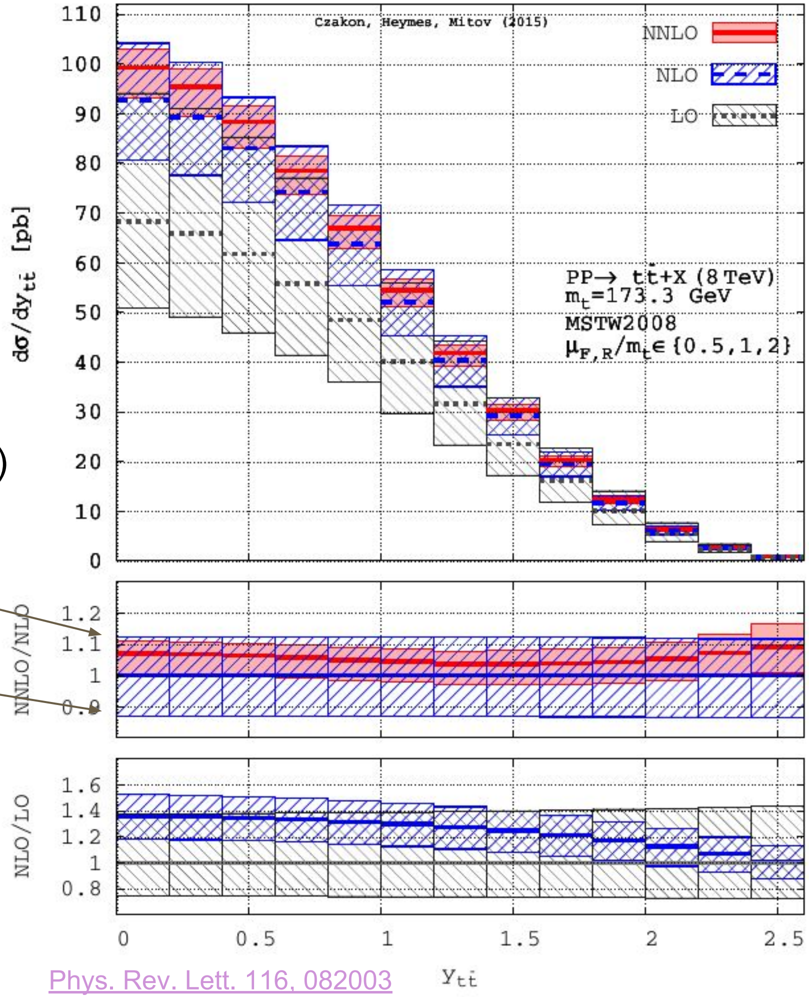
- ***A not-so-obviously tricky case:***
  - scale uncertainties

Take NLO scale variations
as uncertainty (missing NNLO MC)

⇒ flat uncertainty here, and NNLO is within uncertainty, but NNLO/NLO is not flat!

+1

-1

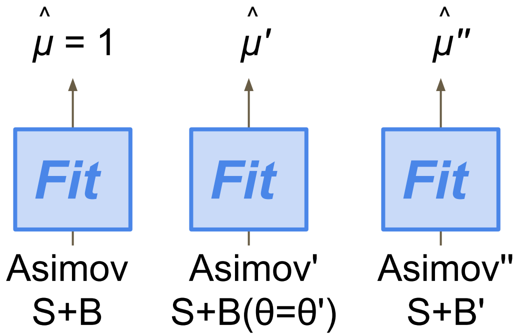


Phys. Rev. Lett. 116, 082003

Suppose data looks like NNLO, we measure y_{tt},
we constrain scale syst. in low y_{tt} bins
⇒ new physics at high y_{tt}?

Systematics model validation

- Especially given the **impossibility to build a "perfect model"**, need to **validate flexibility** of the adopted systematics model *(better if before looking at the full data!)*
- **Response** of the model **to injection tests:**
  - build **toy data** (or *Asimov* data) with **non-nominal properties**
    - can vary **parameters of the model** (*i.e. by shifting NPs when creating the Asimov data-set*)
    - can use a **MC generator not included** in the systematics model to build the toy data
    - check **compatibility** of best-fit POI with injected value



- **Post-fit plots:**
  - evaluate **data/prediction agreement** across distributions with **fit result projection** *(shifted prediction + reduced systematics band)*
  - can spot issues e.g. if found disagreement in a region where don't expect signal
  - especially useful for **validation regions / validation distributions** i.e. regions / distributions not directly used in the fit

Statistical fluctuations on systematics

- Often systematic uncertainty definition affected by **statistical uncertainty**
  - typical example again from **two-point systematics**, evaluated as the difference between **two statistically independent MC samples**, and at least one of them statistically limited

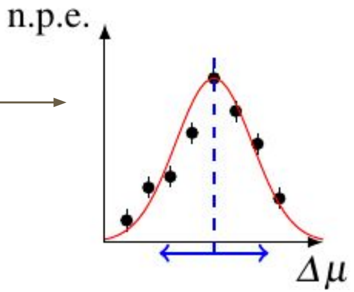
- Current PLR formalism **doesn't account** for this effect:

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\theta}^0 | \mu, \boldsymbol{\theta}) = \prod_{i \in bins} \mathcal{P}(n_i | \mu \cdot S_i(\boldsymbol{\theta}) + B_i(\boldsymbol{\theta})) \times \prod_{j \in syst} \mathcal{G}(\theta_j^0 | \theta_j, \Delta\theta_j)$$

effect of $\pm\Delta\theta_j$ on ***S*** and ***B*** assumed to be **known with no uncertainty**

  - ongoing efforts (in ATLAS at least) to try to incorporate this, e.g. adding NPs for size of ΔB_{ij} = [$B_i(\theta_j=1)$ - $B_i(\theta_j=0)$] $\rightarrow$ $\Delta B_{ij}^0 \pm \Delta(\Delta B_{ij}) = \Delta B_{ij}^0 + \nu_i$ ... → new NPs

- To assess the **size of the effect**:
  - can use **pseudo-experiments / toys** (*e.g. with bootstrap method*)
    - repeat measurement *N* times and see distribution of $\Delta\mu$



- To **mitigate** the effect:
  - ATLAS uses **smoothing of systematic** variations in single distributions
    - assumption of "regular" shape of systematic variation

Including data-driven estimates into the fit model

- Often **data-driven estimates** provided as **inputs** for the PLR model
  - however, ideally better to include the estimate **in the fit model**
    - easier handling of ***correlations***
    - natural way of considering ***signal contamination*** in control region
  - not always possible/easy
    (example: Matrix Method for fakes and non-prompt lepton background determination)

- With more and more data, **natural to consider more** and more
  data-driven or partially data-driven background estimation techniques,
  also for backgrounds currently estimated through MC and with NPs constrained by the data
  - by building more data-driven predictions, possibly included in the PLR model,
    could **reduce issues** related to **over-constraints of NPs**
    and **enhance** **model flexibility**

Assessing "goodness of fit"

- Quantification of **overall goodness** of a PLR fit (like **χ^2-probability**)
  - this ***goodness of fit*** should consider both:
    - data/prediction **agreement** after the fit
    - **pulls** of nuisance parameters
  - can use maximum likelihood value, compared to a reference⇒ ***profile likelihood ratio***
    e.g. with ***saturated model*** (http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf) → used by CMS ttH(bb)

$$\lambda = \frac{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}})}{\mathcal{L}_{\text{saturated}}}$$

built setting prediction = data

g.o.f. = Prob(-2logλ)

- Similar but different problem is to quantify the **post-fit data/prediction agreement** for individual plots/distributions *(which can be part of the fit inputs or **validation** ones)*
  - can define a **χ^2-like variable**, but need to take into account **correlations** between systematics in different bins and correlation between systematics
    - considered in internal review of ATLAS ttH(bb):

$$\chi^2 = \sum_i \sum_j (d_i - x_i) C_{ij}^{-1} (d_j - x_j)$$

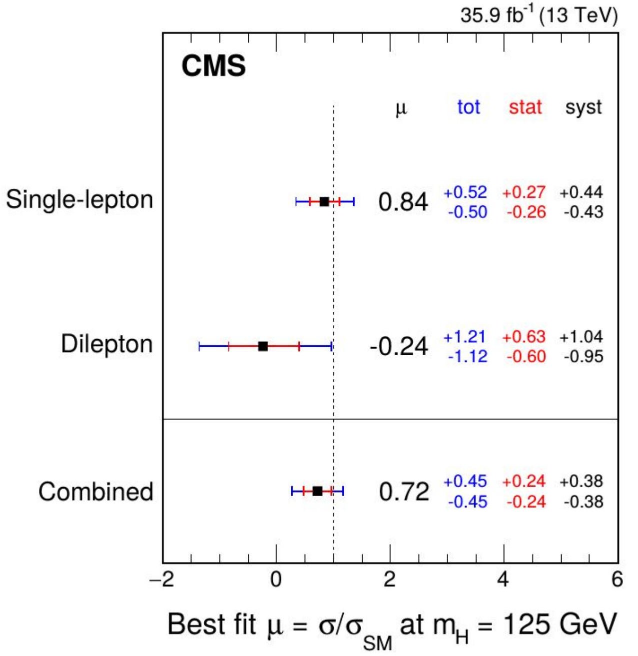
d_i: data in bin *i*; x_i: nominal prediction in bin i

$$C_{ij} = \sum_n \sum_m (x_{i,n} - x_i) \rho_{nm} (x_{j,m} - x_j) + \delta_{ij} x_i$$

$x_{i,n}$: pred. in bin i with syst. variation n; ρ_{nm}: correlation coefficient

Combination of measurements

- With the PLR approach, **combination** of different measurements is **natural**:
  - "just" ***add some more bins*** to the product
- However, important to consider **compatibility of models**:
  - **orthogonality** of channels:
    - bin contents in PLR are supposed to be statistically independent
  - **same** definition of (set of) **POI**:
    - sometimes obvious, but not always
      (is μ applied to all the ttH, or just one decay channel? What about tH? ...)
  - **compatible** set of **systematics**:
    - this is the most tricky part, especially for **ATLAS+CMS combinations**!
    - mainly dealing with the question "***which NPs are correlated between channels?***"
    - often cannot reach perfect solution, need to **test different correlation assumptions**
      (notice that in PLR formalism systematics are either fully correlated or fully uncorrelated...)



Summary and conclusions

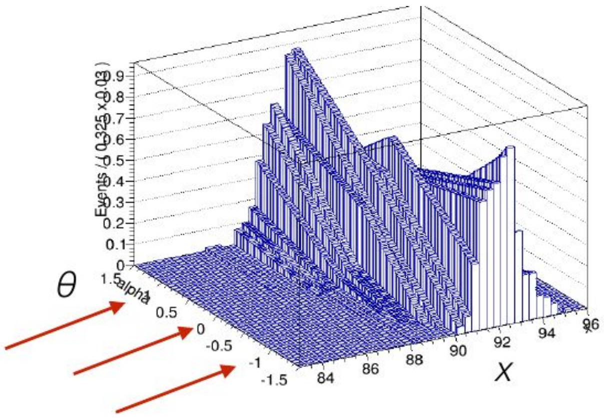
- **Profile likelihood ratio** (PLR) approach presented, in the context of ttH analyses *(mainly in the "binned case", i.e. ttH(bb) and multi-lepton channels)*
  - tod, mefeatures and possible pitfalls discussed

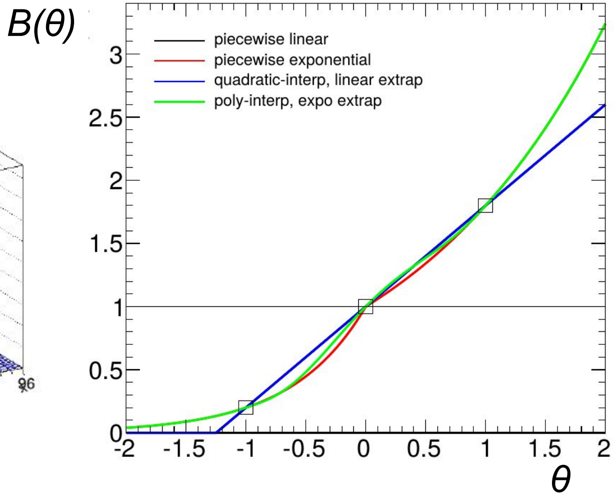
- Room for **fruitful discussion** on the most critical points in these days
  - important to consider **how** the current approach **will work** and eventually **evolve** with:
    - more and more **data**
    - more and more **precision** measurements
    - more and more accurate/refined/rich theory **predictions**
    - new experimental analysis **techniques**...

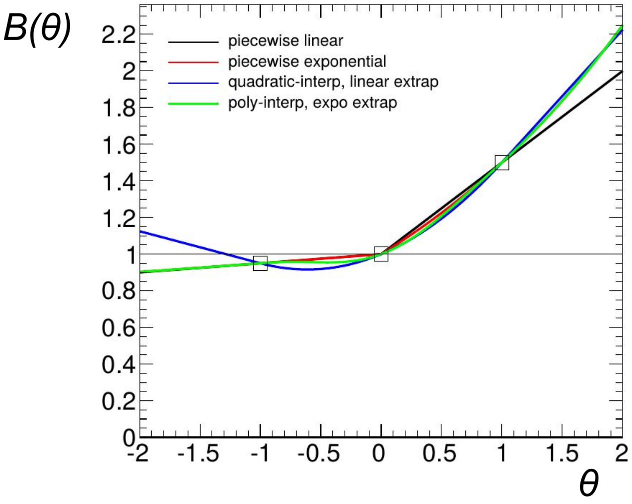
Backup

NP interpolation / extrapolation

- The "**NP interpolation / extrapolation**" controls *how* a **variation of a NP** θ_j reflects in a **variation of the predicted yields** ($\boldsymbol{S}(\theta_j)$ and $\boldsymbol{B}(\theta_j)$, in the various bins)
- Different schemes / codes are implemented in **RooStats**/**RooFit**, e.g.:
  - *piecewise linear* → default for *shape* component of systematics
  - *polynomial interpolation / exponential extrapolation* → default for *norm.* component

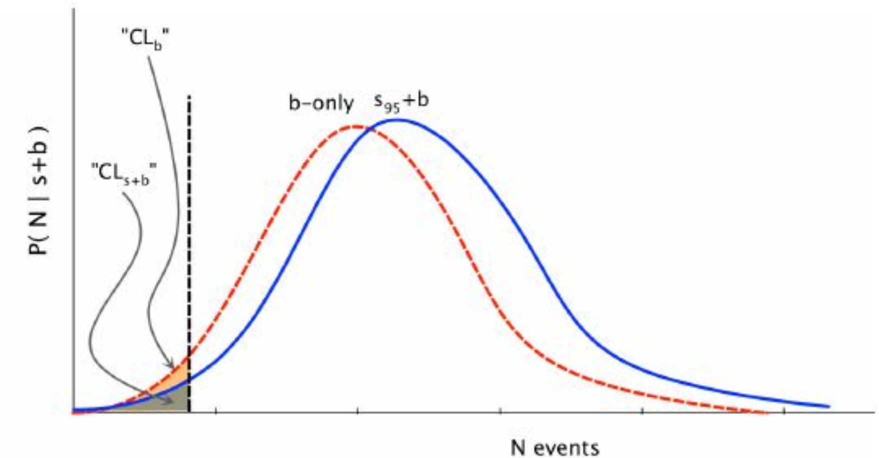






Exclusion limits

- When looking for a tiny signal on top of background, worry to exclude signal due to a downward fluctuation



Using CLs+b, one would expect to exclude the signal 5 % of the time

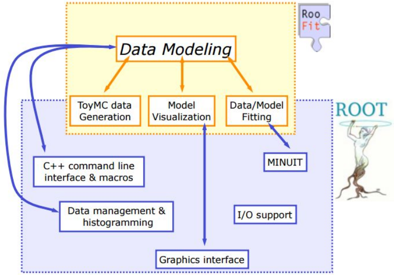
$$CL_s = CL_{s+b}/CL_b$$

test signal hypothesis:

only exclude if $CL_s < 5$ %

- robability):
  - a downward fluctuation in *S+B* will not exclude signal since **CL_b** will also be small

Implementation (in ROOT)



Mathematical concept		RooFit class
variable	x	RooRealVar
function	$f(x)$	RooAbsReal
PDF	$f(x)$	RooAbsPdf
space point	$\vec{x}$	RooArgSet
integral	$\int_{x_{min}}^{x_{max}} f(x)dx$	RooRealIntegral
list of space points		RooAbsData

- **RooFit:** toolkit to extend **ROOT** providing language to describe data models
  - model distribution of observable *x* in terms of parameters *θ* using probability density function PDF

- **RooStats:** project to provide advanced stat. techniques for LHC collaborations
  - built on top of **RooFit**

- **RooWorkspace:** generic container class for all **RooFit** objects, containing:
  - full model configuration (i.e. all information to run statistical calculations)
  - PDF and parameter/observables descriptions uncertainty/shape of nuisance parameters
  - (multiple) data sets

- **HistFactory:** tool for creating **RooFit** workspaces formatted for use with **RooStats** tools
  - meant for analyses based on template histograms