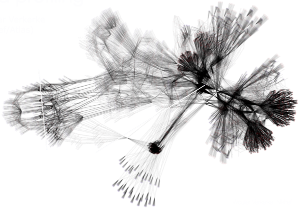
HIP Seminar

Profile Likelihood Methods in Top Quark Mass Measurements at the CMS

Hannu Siikonen



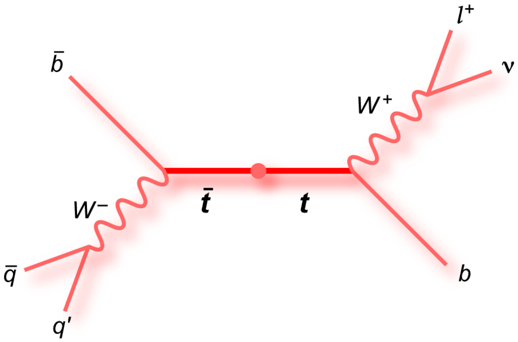
Helsinki Institute of Physics

25th October, 2022

Outline

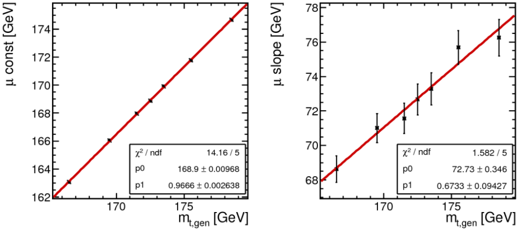
Top Quark Measurements at the LHC

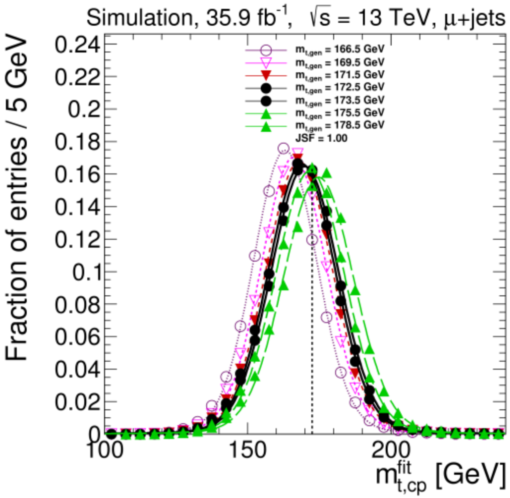
- At the LHC, direct measurements of the top quark mass (m_t) are the most precise
  - Direct ≡ reconstruction of decay products
  - Indirect ≡ anything else, e.g. based on cross-section
- In Run 2 conditions ($\sqrt{s} = 13\,\text{TeV}$):
  - Collision events producing a top quark pair are the most frequent ($832\,\text{fb}^{-1}$)
  - Events with a single top quark make a good number 2 ($264\,\text{fb}^{-1}$)



Visualized: the semileptonic top quark pair decay channel

Direct Top Quark Mass Measurements

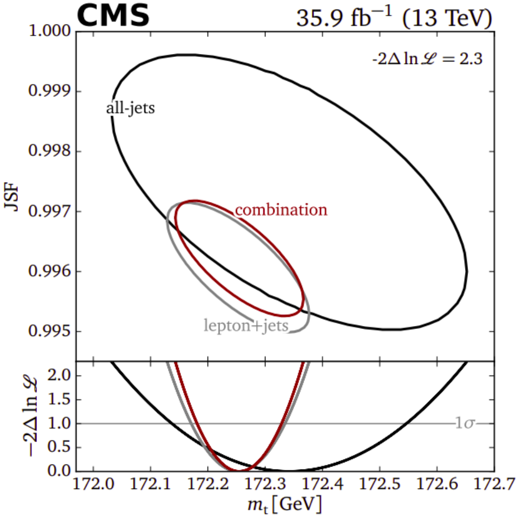




- Classical strategy:
  - Intricate parametrized fits on the observables are made against simulation truth values
  - The parameter dependence on e.g. m_t is fit against the **simulation truth**

Constructing a Likelihood

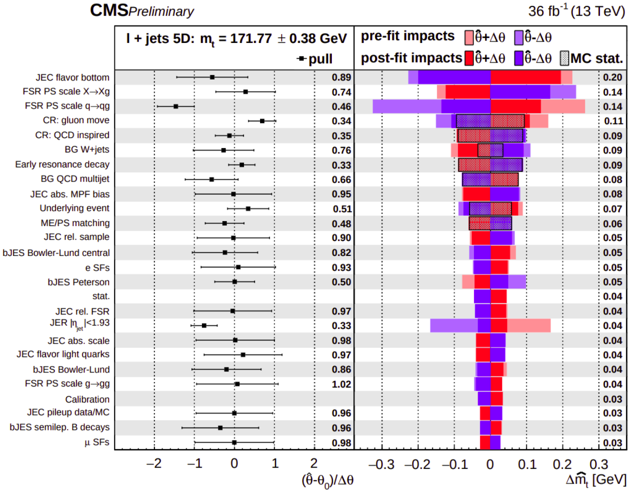
- Using the parametrized fits, a likelihood function for the observables can be constructed
  - This will depend on the chosen parameters
- CMS used to utilize m_t and Jet Scale Factor (JSF)
  - On the right is the **CMS lepton+jets and all-jets likelihood combination on the 2016 data**
  - The yielded result: $m_t = 172.26 \pm 0.07\,(\text{stat+JSF}) \pm 0.61\,(\text{syst})\,GeV$
- ATLAS used in addition bJSF during Run1



Has the Precision Limit been reached?

- The stat. (+ JSF) errors in the CMS 2016 measurement are vanishing vs. the systematics: **0.07 GeV vs. 0.61 GeV**
  - **In the classic treatment, the most important systematic error sources cannot typically be reduced by adding statistics**
  - Such error sources include e.g. modelling uncertainties (in the simulations)
  - This indicates that adding limitless statistics to the measurement would at best yield an error of $\pm$**0.00 (stat+JSF)** $\pm$ **0.61 (syst) GeV**
- As modelling advances, an inverse trend is observed:
  - With better modelling, the number of potential systematic error sources tends to increase
  - With traditional methods, there is always a risk of double-counting and adding statistical noise to each additional error source

Apparently Not?

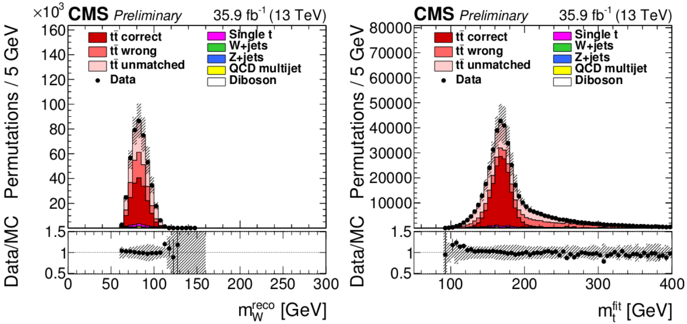


- Earlier this year preliminary results on a new measurement on the 2016 data was released:
  - $m_t = \mathbf{171.77} \pm 0.04\,(\text{stat}) \pm \mathbf{0.38}\,(\text{syst})$ GeV
  - The analysis uses a split scheme of Final State Radiation (FSR) uncertainties: light quarks and heavy quarks handled separately
  - With the old FSR definitions one measures $m_t = \mathbf{172.14} \pm 0.04\,(\text{stat}) \pm \mathbf{0.31}\,(\text{syst})$ GeV

What Has Changed?

- Earlier CMS measurements only used two observables, which excelled in a fit on m_t^{gen} and JSF
  1. The reconstructed m_t resonance
  2. The reconstructed hadronic m_W resonance
- Three new observables were introduced in the new study
  - These include a ratio between b-jet and W jet p_T values, earlier utilized by ATLAS
  - Most importantly, the use of the partial resonance between a b-jet and the charged lepton was added in **a phase-space region that was earlier unused**
- **These updates are still not sufficient for explaining the improvement**
  - The missing piece is the introduction of a profile likelihood approach

A "Simple" In-Situ Measurement: JSF



- In the earlier iterations, the JSF **nuisance parameter** is measured in association to m_t in an *in-situ* manner
  - Measuring the additional dependence is possible, as the m_W distribution is present in the full likelihood
  - This allows reducing the jet calibration systematics

More Nuisance Parameters

- **Could we do the same thing more comprehensively, making nuisance parameters from all systematic error sources?**

A Generalized In-Situ Measurement

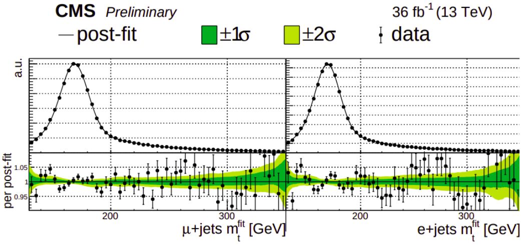
- There is no fundamental obstacle for interpreting the systematic variations similarly as JSF variations
  - **This is exactly what is done in the new CMS m_t analysis on the 2016 Data with a profile likelihood method**
  - In practice this requires a higher level of automation
  - In a final analysis there can easily exist more than 100 systematic uncertainties
- In consequence, the auxiliary JSF parameter becomes unnecessary:
  - This task is already covered by the Jet Energy Uncertainties
- Systematic uncertainties are in general modelled with variations on the central simulated samples
  - This includes two-sided variations, interpreted as the corresponding $\pm\sigma$ uncertainties
  - …and one-sided variations, which are interpreted as $+\sigma$ uncertainties

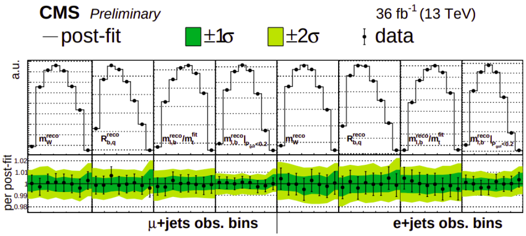
Systematics in the Likelihood

- In the old CMS analyses the best results were found with the **hybrid method**:
  - Here, a physically motivated Gaussian prior was imposed on the value of JSF
- In the profile likelihood approach the systematics are interpreted as nuisance parameters θ_k
  - The $\pm\sigma$ variations are defined to correspond to the values $\theta_k = \pm 1$
  - The knowledge that the variations are the $\pm\sigma$ uncertainties is enforced by a Gaussian prior $\mathcal{G}$ on all of the nuisance parameters θ_k
  - Similar Gaussian priors are also set on the normalization scales (η_j) of the simulated samples, according to the cross-section and luminosity uncertainties
- In summary, starting from the likelihood without priors $(\mathcal{L}_0)$, the full likelihood stands as:

$$\mathcal{L} = \mathcal{L}_0 \times \prod_{k \in \text{nuisances}} \mathcal{G}(\theta_k) \times \prod_{j \in \text{samples}} \mathcal{G}(\eta_j) \,. \tag{1}$$

Parameter Automatization?





- The automatization ends up being a major issue in the practical implementation
  - **One method for tackling the issue is using a binned likelihood model**
  - Here, the only parameters are the bin edges, and the only variables to fit the bin event counts
  - Each bin forms its own Poisson counting experiment
- The 2016 CMS m_t analysis started with the 5 observables being handled by fit functions, but ended up handling 4/5 observables with binning (left)

Considerations for a Binned Analysis

- Maybe counter-intuitively, the binned approach is the most robust choice
  - **Parametrized fit functions are notoriously volatile, even with a very good function**
  - In contrast, a binned analysis derives its robustness from the simplicity of Poisson statistics
- The CMS 2017–2018 m_t analysis uses the same 5 observables as the 2016 analysis, but all with a binned approach
  - This will be further handled in my **thesis defence in Chemicum A129 on 22nd November, 1 pm**

A Binned Likelihood

- For a fully binned analysis the core likelihood in Eq. (1) can be expressed as:

$$\mathcal{L}_0 = \prod_{i \in \text{bins}} \mathcal{P}\left(n_i \middle| \sum_{j \in \text{samples}} (1+\kappa_j)^{\eta_j} \times \nu_i^j\left(\vec{\theta}, m_t\right)\right) \quad (2)$$

  - $\mathcal{P}(n_i|\lambda)$ is the Poisson probability distribution for the (bin) event yield n_i
  - ν_i^j the expected event yield for the simulated sample j in the bin i
  - $\vec{\theta}$ collects the nuisance parameters
  - κ_j is the fraction of normalization uncertainty for the sample j, controlled by the nuisance parameter η_j.

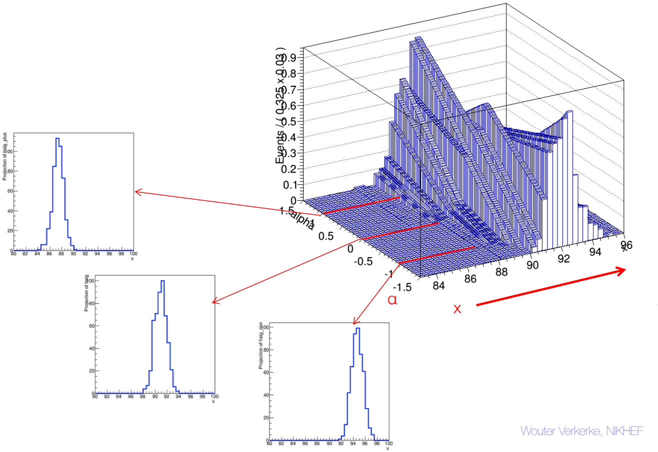
- As a reminder, Eq. (1) stands as

$$\mathcal{L} = \mathcal{L}_0 \times \prod_{k \in \text{nuisances}} \mathcal{G}\left(\theta_k\right) \times \prod_{j \in \text{samples}} \mathcal{G}\left(\eta_j\right).$$

Interpolation?

- For a binned approach, interpolation between the central simulation sample and the variations is achieved through **histogram morphing techniques**

Visualization of bin-by-bin linear interpolation of distribution



Wouter Verkerke, NIKHEF

Morphing Strategies

- The two main classes of histogram morphing are vertical morphing (see previous slide) and horizontal morphing
  - Vertical morphing (seen on the previous slide) is relatively simple: each bin is considered separately
  - Horizontal morphing can be implemented with various algorithms, and it usually considers both bin migration and vertical effects
- Vertical morphing is preferred by its simplicity:
  - No migration between bins, so each bin receives its own parameters
    - The morphing parameters are easy to determine automatically
    - Typically a smooth interpolation function is utilized
  - The parametrization of each nuisance for each bin can be parametrized either ...
    - as a multiplicative factor (more common) or ...
    - as a direct offset

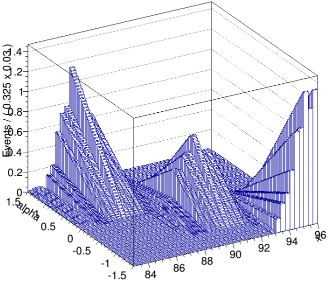
Vertical vs. Horizontal Morphing

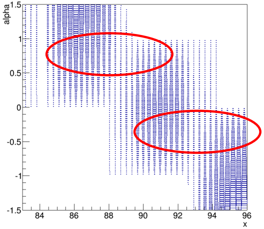
- Even if vertical morphing is preferable, horizontal morphing becomes necessary when the variations are too big

Limitations of piece-wise linear interpolation

- Bin-by-bin interpolation looks spectacularly easy and simple, but be aware of its limitations
  - Same example, but with larger 'mean shift' between templates

Note double peak structure around |α|=0.5





Wouter Verkerke, NIKHEF

Determination of a Good Binning

- Strict rules follow from the preference of vertical morphing:
  - **Bins must be wide enough**, so that the variations impose relatively small changes in bin contents
  - Simultaneously the bins should be **narrow enough** to provide resolution to measure the problem at hand
- Different problems may prefer different solutions:
  - A "bump hunt" is best done on an evenly spaced binning on the mass axis
  - **A precision measurement (such as that of m_t) is best performed with bins with even statistics**

What is This All Good For?

- Nomenclature:
  - There can be one or more **unconstrained Parameters of Interest**
  - For us, there is only one: $\mathbf{m_t}$
  - This is in contrast to the nuisance parameters $\vec{\theta}$ with Gaussian constraints
- We can define the following **profile likelihood ratio**:

$$\lambda(m_t) = \frac{\mathcal{L}\left(m_t, \hat{\hat{\vec{\theta}}}_{m_t}\right)}{\mathcal{L}\left(\hat{m}_t, \hat{\vec{\theta}}\right)} \tag{3}$$

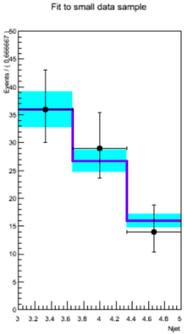
  - Here, $\hat{\vec{\theta}}$ is the global maximum likelihood solution for the nuisance parameters
  - And $\hat{\hat{\vec{\theta}}}_{m_t}$ the solution at m_t
- According to Wilks' theorem $-2\ln\lambda(m_t)$ asymptotically approaches the χ^2 distribution at good statistics (error $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ for the 1D profile)

Nuisance Parameter Constraints

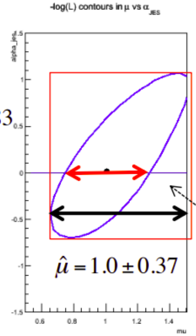
- The *a priori* $\pm\sigma$ uncertainties correspond to the index k nuisance parameter values $\theta_k = \pm 1$
- Considering the full likelihood function, **the nuisance parameter can be constrained *a posteriori* to a smaller uncertainty**
  - This can be understood through applying Wilks' theorem to the nuisance
  - The nuisance parameter $\pm\sigma$ (*a posteriori*) **limits are found at** $\mathbf{-2}\ln\lambda\left(\theta_{\mathbf{k}}\right) = \mathbf{1}$
  - Here, $\lambda\left(\theta_k\right)$ is understood in analogy to $\lambda\left(m_t\right)$ – i.e. all parameters but θ_k being fixed at the maximum likelihood values
- On the next two slides an example borrowed from Ref. 4 is presented

Example: Loose Constraint

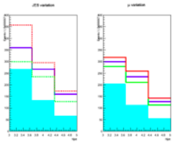
Now measure (μ,α) from data – 80 events



$\hat{\alpha} = 0.01 \pm 0.83$



$\hat{\mu} = 1.0 \pm 0.37$



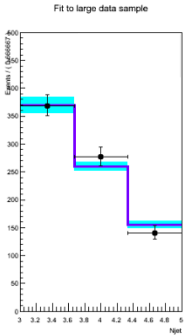
Estimators of μ, α correlated due to similar response in physics measurement

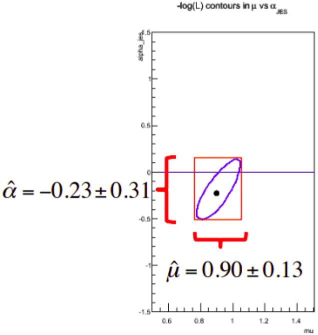
Uncertainty on μ with/without effect of JES

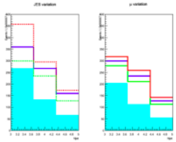
- The Parameter of Interest μ is measured aside the Jet Energy Scale (JES) nuisance α, yielding a slight constraint on α from the *a priori* values $\alpha = \pm 1$

Example: Tight Constraint

The next year – 10x more data (800 events)
repeat measurement with same model







Estimators of μ, α correlated due to similar response in physics measurement

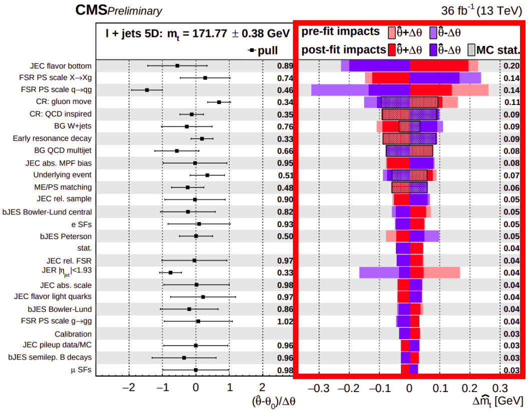
- Here, statistics start to take over, constraining α to ± 0.31, bringing down also the uncertainty on μ

Which Nuisances can be Constrained

- **It is important to emphasize that the systematics that can be constrained usually have an experimental connection**
  - This includes e.g. the energy scales of jets and charged leptons
  - . . . and jet flavor tagging (mainly b-jets)
- But also the choices in the simulation are uncertain and (maybe less obviously) connected to the experiment:
  - Parton Showers (Initial and Final State Radiation)
  - Parton Density Functions
  - Hadronization
  - Underlying event

Nuisance Parameter Impacts

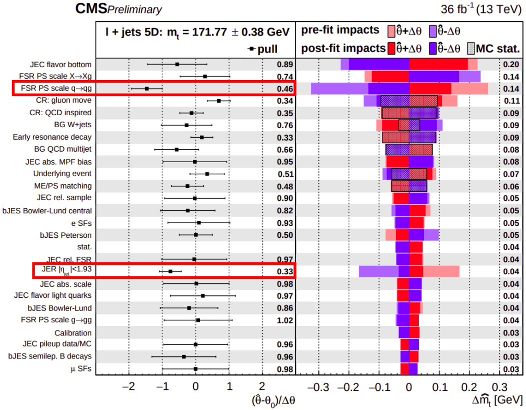
- The uncertainty δm_t imposed by the index k nuisance θ_k on the parameter of interest (m_t) is called the **impact** of θ_k on m_t
- The impacts are found in a few steps
  - First fitting all the nuisance parameters and the parameter of interest
  - Then, fixing θ_k to $+1$ and -1, and then fitting on all the other parameters
  - The difference in the value of the parameter of interest is the impact



- **The same approach works both on the *a priori* nuisance limits $\theta_k = \pm 1$ and the possibly constrained *a posteriori* limits $\theta_k = \pm\sigma_{\text{posteriori}}$, where $|\sigma_{\text{posteriori}}| < 1$**

Nuisance Parameter Pulls

- A **pull** is defined as the nuisance parameter offset from zero (expressed in the units of the *a priori* $\pm\sigma$ uncertainty)
- Impacts an pulls can be **blinded or unblinded**:
  - In the blinded mode simulation is used instead of real data
  - The simulation agrees with the central values in the model, so all blinded pulls are zero
  - **However, the impacts are meaningful estimators for the true errors that will be measured in data**

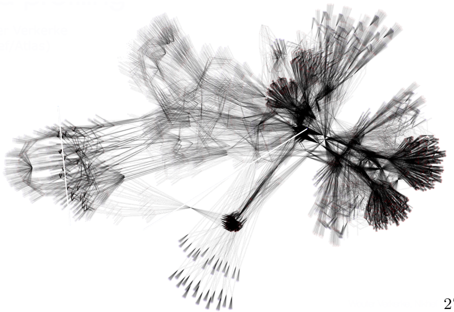


- In the unblinded case notable pulls can appear but they are expected to be less than a unity (one sigma)

Statistical Uncertainties

- When estimating the expected number of events in a bin, the **statistical uncertainties are also a kind of modelling uncertainty:**
  - This means that every simulated sample in each bin should receive their statistical nuisance parameter
  - At low statistics they follow the Poisson distribution, which converges to a Gaussian at large statistics
  - If around 10 simulated samples are combined, the number of parameters quickly explodes

- Visualized from Ref. 4: Atlas Higgs combination model (23.000 functions, 1600 parameters)
- Partial cure: the **Barlow-Beeston** approach allows the combination of statistical errors from separate samples

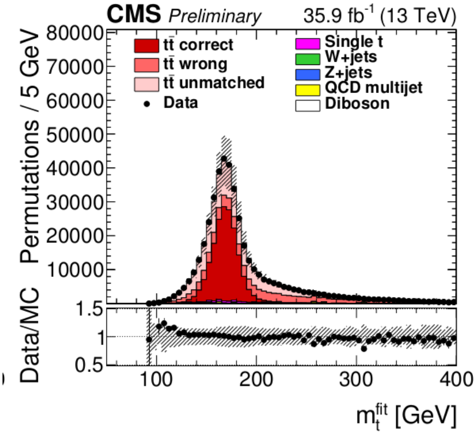


Combination of Simulations

- Integrated luminosity $\mathcal{L}$ is the general measure for the amount of events in a sample
- When simulation and data are compared, each sample should be weighted by:

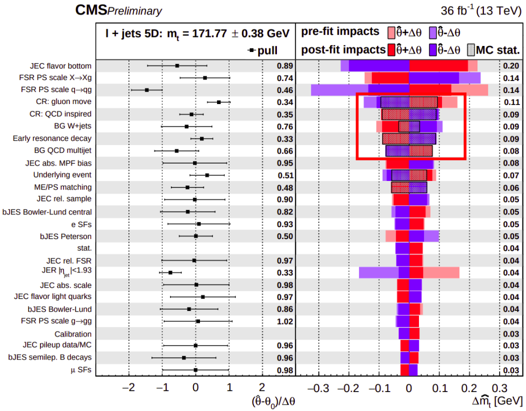
$$w^{\mathrm{Sim}} = \frac{\mathcal{L}^{\mathrm{Data}}}{\mathcal{L}^{\mathrm{Sim}}_{\mathrm{Eff}}} = \frac{\sigma^{\mathrm{Sim}}\mathcal{L}^{\mathrm{Data}}}{N^{\mathrm{Sim}}_{\mathrm{Eff}}} \tag{4}$$

- σ^{Sim} is the cross-section of each simulated sample (Sim)
- $N^{\mathrm{Sim}}_{\mathrm{Eff}}$ is the effective number of events in each simulated sample, considering event weights



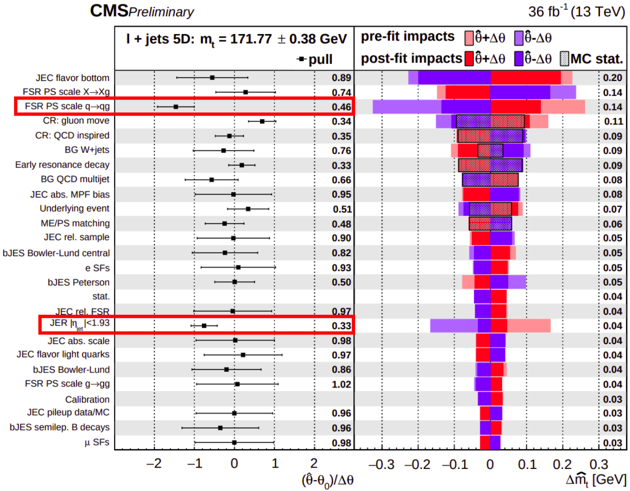
Statistical Variation Uncertainties

- There are three common methods for quantifying systematics
  1. Reweighting the simulated sample (almost 100% statistical correlation)
  2. Rescaling object (e.g. jet) energies (quite close to 100% statistical correlation)
  3. Separate simulated samples (no statistical correlation)



- In the third case one should provide separate bin-wise nuisance parameters for the systematic variation sample
  - **This is often a weak point and poorly implemented in the common tools**

A Final Look at the New CMS Results



- In the new 2016 analysis especial quark Final State Radiation (FSR) and Jet Energy Resolution (JER) are constrained to less than ± 0.5
- This is explained by the hadronic W boson resonance
- The constraints are OK, but the notable **pulls** are a point of concern
- Without the constraints, the total error would be around ± 0.5 GeV

What to expect for 2017–2018

- **The number of recorded events is around 3 times larger in 2017–2018 w.r.t. 2016**
  - Statistical error for N events is according to Poisson statistics proportional to $1/\sqrt{N}$
  - For systematics that are well constrained by the measured data, one can optimistically expect similar scaling as for statistical errors
  - In consequence, the error scaling factor in 2017–2018 w.r.t. 2016 can go to $\mathbf{1/\sqrt{3} \approx 0.58}$
- Other improvements:
  - The method is more stable with a binned with also for m_t
  - The 2016 analysis scales away the even yield both in the central simulation and systematic variations
  - With Poisson statistics the (absolute) event yields are an important part of the whole picture, so **we see including these as an important improvement**

Historical Review

- **One could ask *why now?***
- Tools have evolved and gotten more wide-spread
- Slow diffusion from one field of study to another:
  - The early adopters at the LHC were the Higgs hunters, as can be reviewed from this profile likelihood article and this article on CMS/ATLAS methods on Higgs boson searches
  - In the top quark community, the methods first spread to inclusive cross-section measurements, and then to differential cross-section measurements
  - At the final step, also precision measurements (such as the m_t measurement) are taking up the tools

Summary

- **The most precise m_t measures at the LHC are currently performed on the top quark pair topology, based directly on the top quark decay products**
- In early Run 2 it seemed that the precision limit for such m_t measurements had been reached
  - This is ruled by irreducible (systematic) error sources
- Profile likelihood methods ruled this as a false assumption
  - The systematic errors can be constrained *in-situ* by the measurement
- Interesting future results coming up on the 2017-2018 CMS Run2 data:
  - A first look at these are given in my thesis defence on Tuesday November 22nd 1 pm at Chemicum A129

Sources

- These slides can be reached at tinyurl.com/hiphannu

1. https://cds.cern.ch/record/2674989
2. https://arxiv.org/pdf/1812.10534.pdf
3. https://cds.cern.ch/record/2806509/files/TOP-20-008-pas.pdf
4. https://www.precision.hep.phy.cam.ac.uk/wp-content/people/mitov/lectures/GraduateLectures/Advanced-Statistics-Verkerke.pdf