TRExFitter: A framework for binned template profile likelihood fits in ATLAS

Santu Mondal

New Vistas in Photon Physics in Heavy-Ion Collisions

Institute of Nuclear Physics Polish Academy of Sciences & AGH University of Science and Technology

20.09.2022

Statistical analysis

● Statistical Analysis in particle-collider physics:
The way to extract quantitative information from collision data

In this talk:

- Quick review of basic principles methods for statistical analysis in HEP
- Introduction to the TRExFitter
- Working mechanism & features of TRExFitter
- Current usage in publications
- Summary and overview

TRExFitter

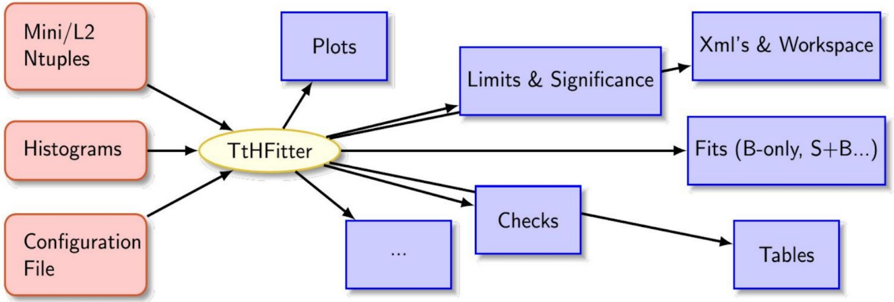
- A profile likelihood fit package, powerful, configurable.
- Born with the name “TtHFitter” (in 2015), as a user-friendly interface to perform fits, extract CLs limits and produce post-fit data-vs-MC plots
- Later became more powerful, changing name to “**TRExFitter**”
- **TRExFitter** based on binned profile likelihood, with statistical inference based on maximum-likelihood principle, profile-likelihood-ratio test-statistics and asymptotic approximation

IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

TRExFitter

- **TRExFitter** is a framework to create and operate statistical models
  - Create RooFit workspaces, through HistFactory
  - Process them through widely used RooStats macros to perform profile-likelihood fits, extract CLs limits and produce post-fit data-vs-MC plots
  - Actively developed and used in many physics analyses : cross-section-fitting / signal-discovery machinery



Maximum likelihood and Fits

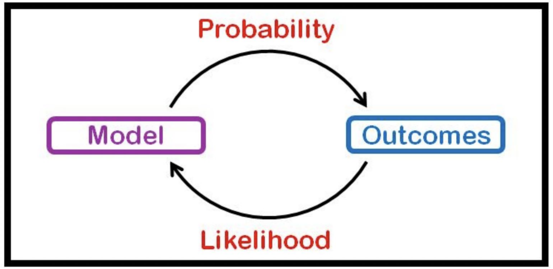
- Likelihood:

  Defined as probability of observing a certain set of data given a model / hypothesis (with certain parameter values)

$$L(\vec{\theta}) = Prob(\vec{x}|\vec{\theta}) = \prod_i Prob(x_i|\vec{\theta})$$

probability *data* *parameters*

if data points / measurements / observation are independent (i.e. uncorrelated)



- Maximum Likelihood principle:

  estimated value(s) of parameter(s) = value(s) maximizing the Likelihood

- “Fit”:

  parameter estimation procedure via Likelihood maximization

IEAP CTU in Prague
SLEZSKÁ UNIVERZITA V OPAVĚ

Profile likelihood

- In the case of a likelihood function depending on many parameters, but where one is interested in only one parameter μ and its uncertainty, one can use a *profile likelihood ratio* defined as

maximum likelihood
for given μ
↓

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

↑
unconditional
maximum likelihood

In the numerator, the parameters **θ** are fitted to their MLE, $\hat{\hat{\boldsymbol{\theta}}}$ for a given value of the parameter μ. In the denominator, μ is also estimated – the values $\hat{\mu}$ and $\hat{\theta}$ define the global maximum of the likelihood *L*.

- This method of profiling the likelihood is very popular for estimating uncertainties from a maximum-likelihood fit; in high energy physics it is known as the *Minos* method of the minuit program

The HistFactory model

- HistFactory is the standard model used in ATLAS for binned statistical analysis

- Specifies how to construct the likelihood function from a set of building blocks

  - Channels (also called regions in **TRExFitter**) are regions of phase space
  - Distributions of samples (MC and data) in channels are provided by template histograms
  - Systematics act on samples and are specified via the distribution at ±1σ shifts

$$p(\vec{n}, \vec{a} \mid \vec{k}, \vec{\theta}) = \prod_i \mathrm{Pois}(n_i \mid \nu_i(\vec{k}, \vec{\theta})) \cdot \prod_j c_j(a_j \mid \theta_j)$$

observed data

unconstrained parameters, e.g. POI

prediction (summed over samples)

constraint term (e.g. Gaussian)

auxiliary data, e.g. from CP group calibration measurement

constrained nuisance parameters

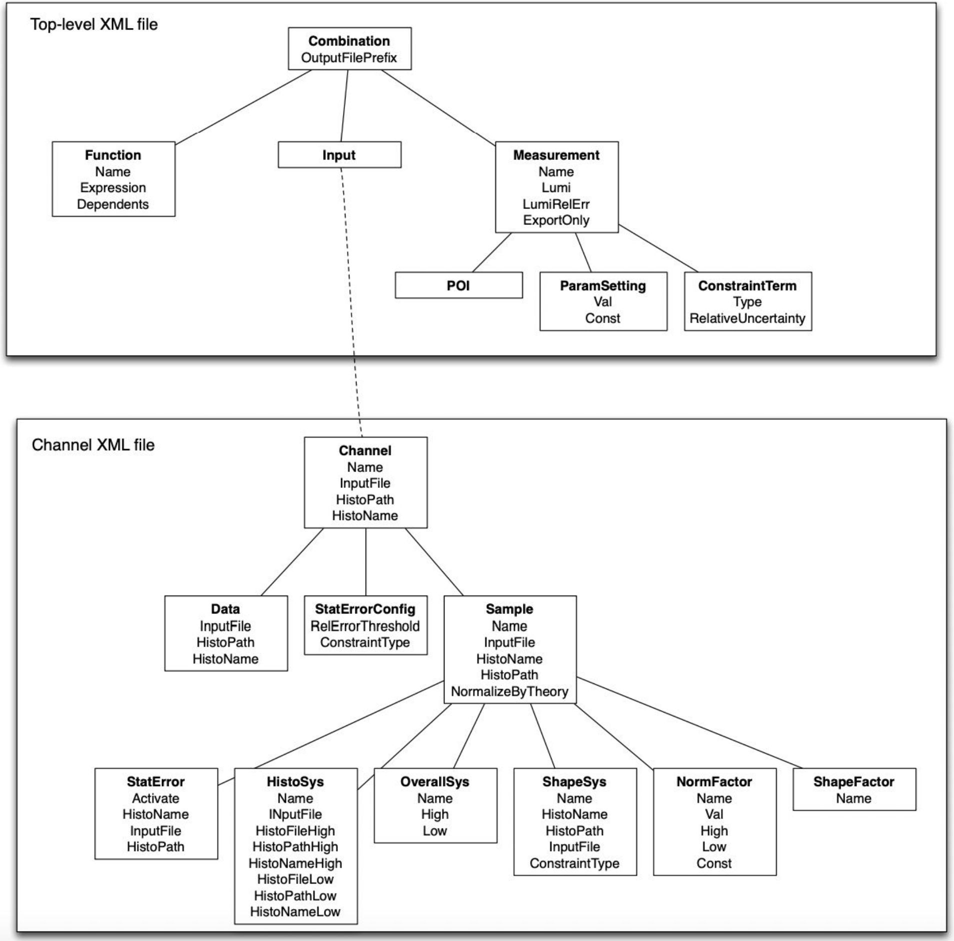
product over all bins in all channels

https://pyhf.github.io/pyhf-tutorial/IntroToHiFa.html

IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

The HistFactory model



Data Analysis in High Energy Physics, A Practical Guide to Statistical Methods -

Olaf Behnke, Kevin Kröninger, Grégory Schott, and Thomas Schörner-Sadenius

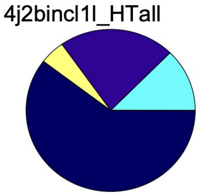
IEAP CTU in Prague

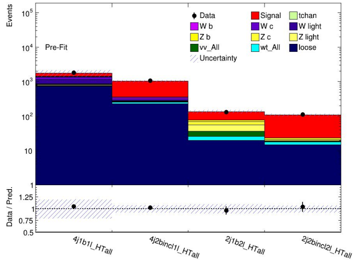
SLEZSKÁ UNIVERZITA V OPAVĚ

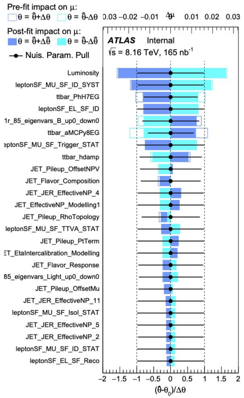
Using TRExFitter

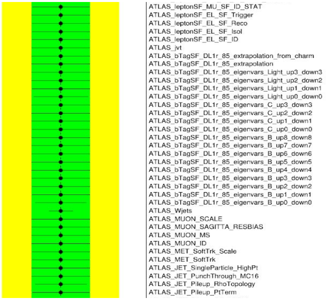
Hang out with TRExFitter

- Declare a fit model, and provide input ntuples or histograms
- Framework provides diagnostics tools and allows to easily adjust the fit model to study the fit
- **TRExFitter** is controlled via a declarative configuration and a command line interface (CLI)
- “Steps” or “actions” in the CLI correspond to tasks executed by **TRExFitter**
- **TRExFitter** produces a lot output :
    - Figures: data/MC, fit model details, statistical inference results, …
    - Tables: yields, effects of systematic ‣ ROOT, txt, YAML files with additional information
    - Also check for warnings and errors in the output!
    - Job settings contain methods to customize output









IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

The configuration file

- Configuration file follows a custom plain text format
- Split into **blocks**, separated by blank lines
  - **Job block:** general options
  - **Fit block:** configuration of fit options
  - **Region blocks:** define distributions included in fit - Exception: validation regions, but can project fit result onto them
  - **Sample blocks:** samples (data + MC) considered in fit
  - **Systematic blocks:** systematic uncertainties affecting samples

```
Job: "FitExample"
  Label: "Fit Example"
  CmeLabel: "13 TeV"
  LumiLabel: "300 fb^{-1}"
  POI: "SigXsecOverSM"
  ReadFrom: HIST
  HistoPath: "ExampleInputs"
  DebugLevel: 2
  SystControlPlots: TRUE
  UseGammaPulls: TRUE

Fit: "myFit"
  FitType: SPLUSB
  FitRegion: CRSR
  doLHscan: SigXsecOverSM

Region: "SR_1"
  Type: CONTROL
  HistoName: "HTj"
  VariableTitle: "H_{T} [GeV]"
  Label: "Signal Region 1"
  ShortLabel: "SR 1"

Sample: "Data"
  Title: "Data 2015"
  Type: data
  HistoFile: "data"

Sample: "Bkg1"
  Type: BACKGROUND
  Title: "Background"
  FillColor: 400
  LineColor: 1
  HistoFile: "bkg1"

Systematic: "JES"
  Title: "Jet Energy Scale"
  Type: HISTO
  HistoNameSufUp: "_jesUp"
%  HistoNameSufDown: "_jesDown"
  Samples: Bkg1,Signal
  Smoothing: 40
%  Symmetrisation: TwoSided
  Symmetrisation: ONESIDED
  Category: Instrumental
```

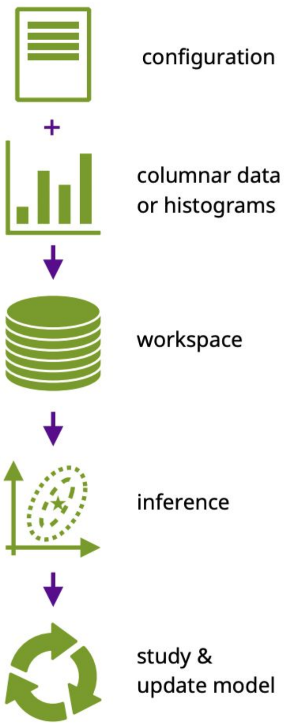
Basic workflow with TRExFitter

Define a fit model in a declarative configuration file:

- **n/h** step: **TRExFitter** reads input (ntuples or histograms) and produces histograms
- **w** step: **TRExFitter** constructs a HistFactory workspace from all template histograms
- **f** step: maximum likelihood fit
- **d/p** step: Pre-/post-fit data/MC visualization
- **r/i** - Nuisance parameter ranking and impact

 step: **TRExFitter** steers statistical inference and visualizes results

- **s** step: discovery significance
- **l** step: parameter limits

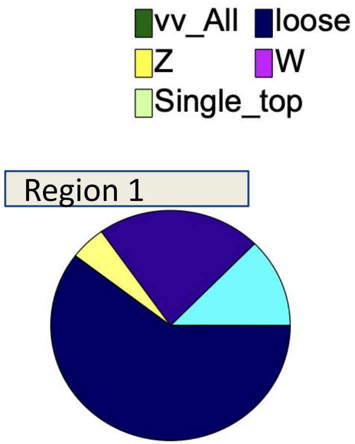


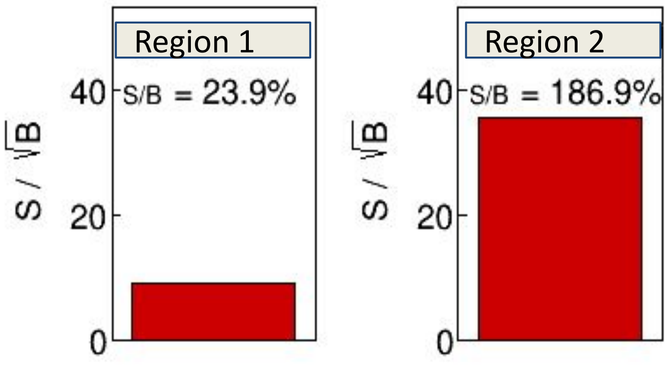
- **TRExFitter** can run multiple regions at the same time.
- Modify fit model, study changes, converge on final model to be used in analysis

IEAP CTU in Prague SLEZSKÁ UNIVERZITA V OPAVĚ

Signal Regions and Compositions

- **PieChart** and **SignalRegions** show background composition and fraction of signal in the regions defined

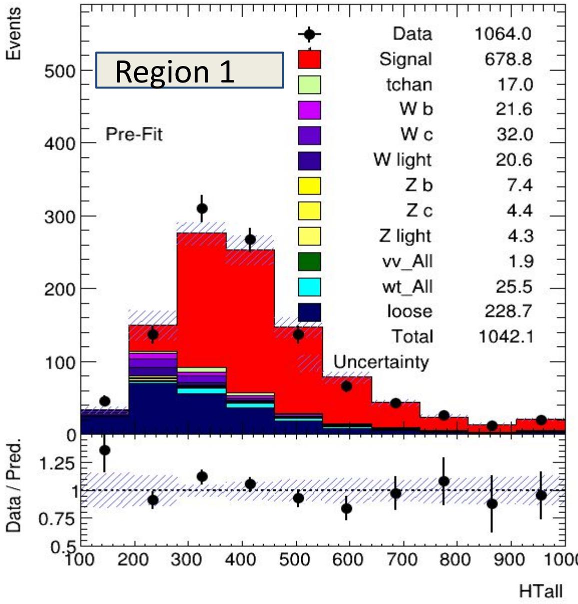




Histogram creation and first plots

- **TRExFitter** supports reading both ntuple and histogram inputs (via **n/h** steps)

  One plot like this generated per analysis region (channel)
  - Total uncertainty of all sources evaluated and visualized
  - Algorithms to automatically obtain suitable binning
  - Especially useful for MVA output distributions
  - Can of course also specify bins by hand

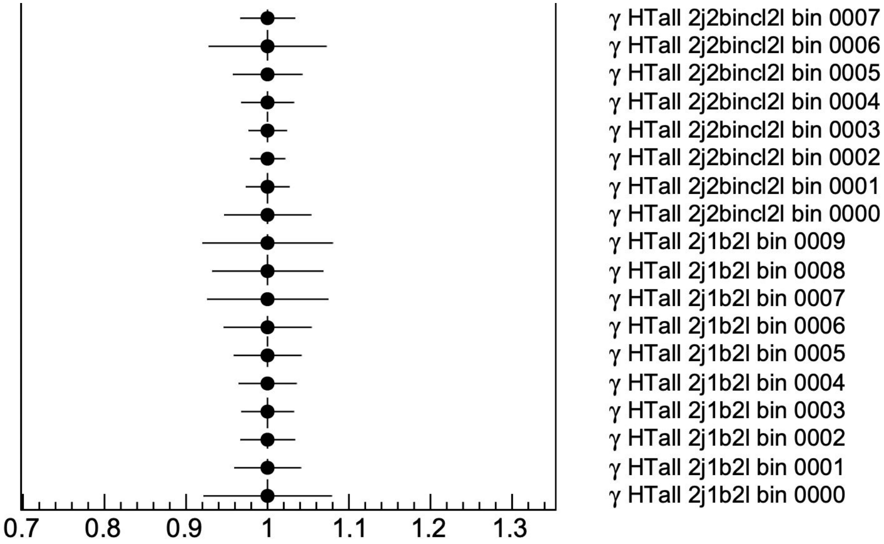


- Pre-fit fit model visualization via **d** step

  - provides data/MC plots and yields per region (channel), summary plots, background composition, S/B, etc.
  - Can customize appearance for publication-quality figures

IEAP CTU in Prague
SLEZSKÁ UNIVERZITA V OPAVĚ

MC statistical uncertainty



Statistical uncertainty in prediction:

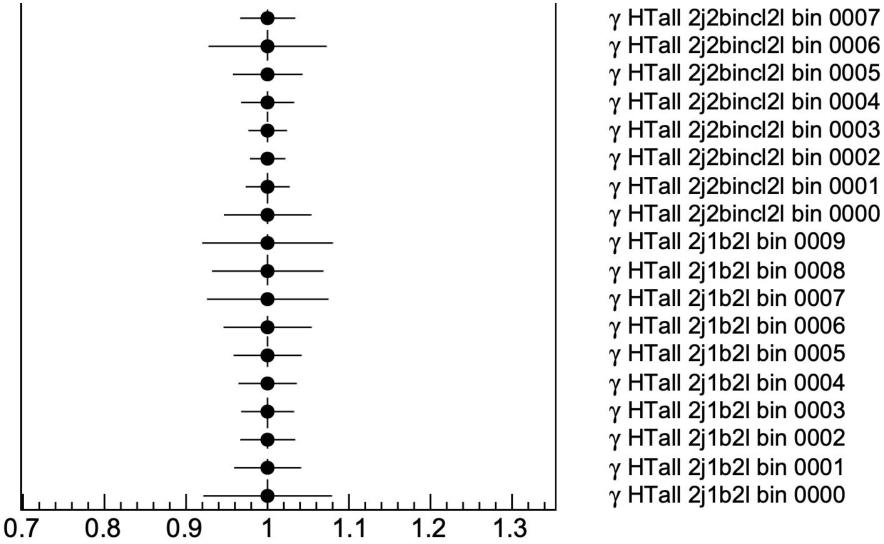
- Model uncertainties due to the finite number of events in simulation are described by dedicated nuisance parameters called **gammas**

- **TRExFitter** automatically creates these parameters for you.

IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

MC statistical uncertainty



Asimov dataset:

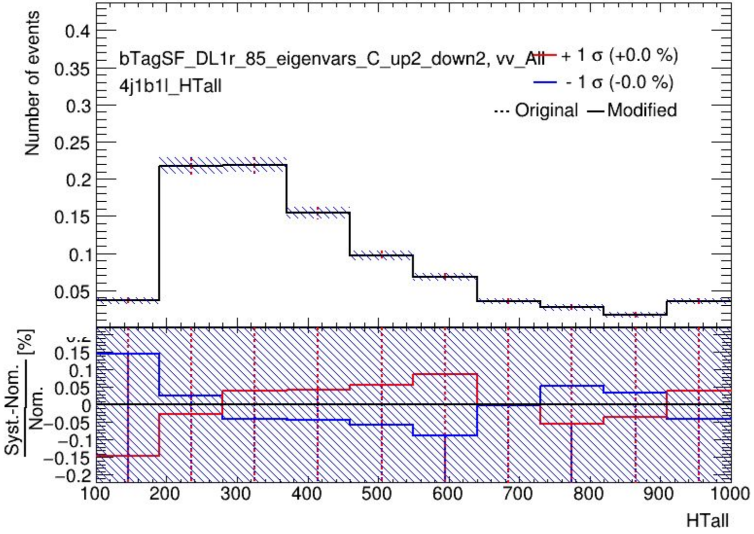
- Dataset for which the maximum likelihood estimate of all parameters matches their “true” value - Can be built without reference to data
- No pulls when fitting this dataset
- Useful for studying expected performance (uncertainties, significance, limits, …)

Investigating systematics templates

- **TRExFitter** visualize the effect of all systematic variations
  - Per Region (channel), per sample, per variation
  - Important to validate the physics

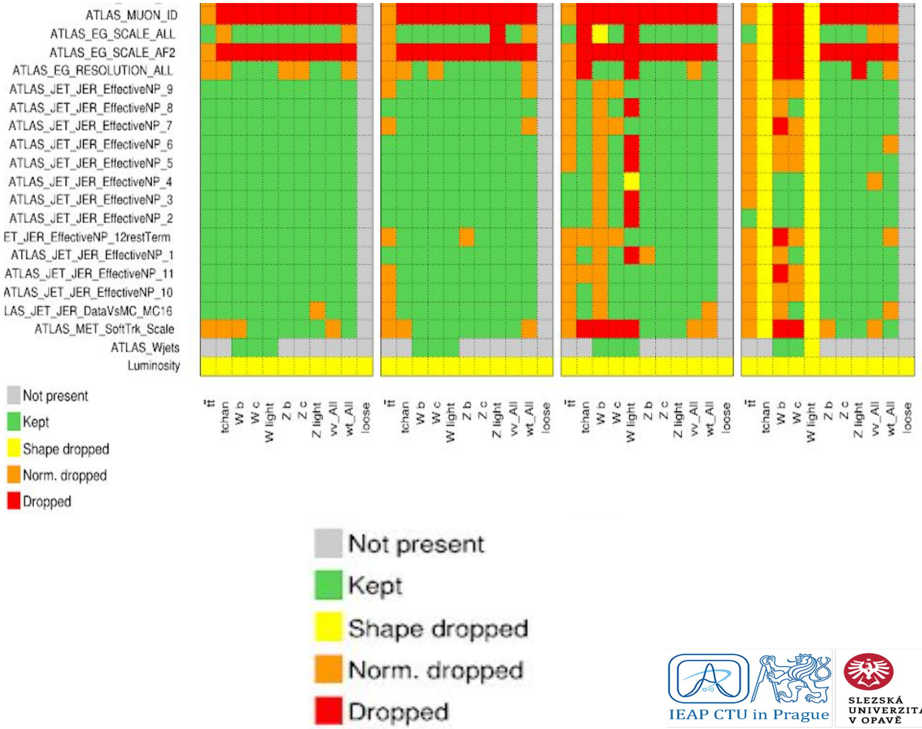
- Study these plots to ensure fit inputs are robust
  - Strange behavior frequently caused by template issues

- Each (independent) source of systematic uncertainty included in the likelihood as constrained NP:
  - Affecting S+B prediction in a coherent way
  - Effect interpolated and extrapolated from 3 discrete values (0 = nominal, 1 = “up” var., -1 = “down” var.) to range of continuous values



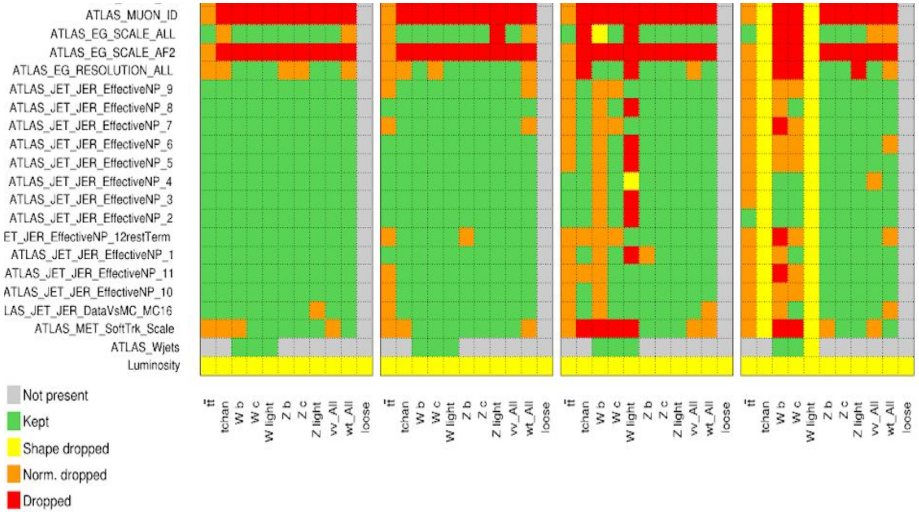
IEAP CTU in Prague
SLEZSKÁ UNIVERZITA V OPAVĚ

Workspace production and pruning



Workspace production and pruning

- Pruning at this step removes negligible effects from systematic variations
  - Control the threshold via
    - SystPruningNorm
    - SystPruningShape

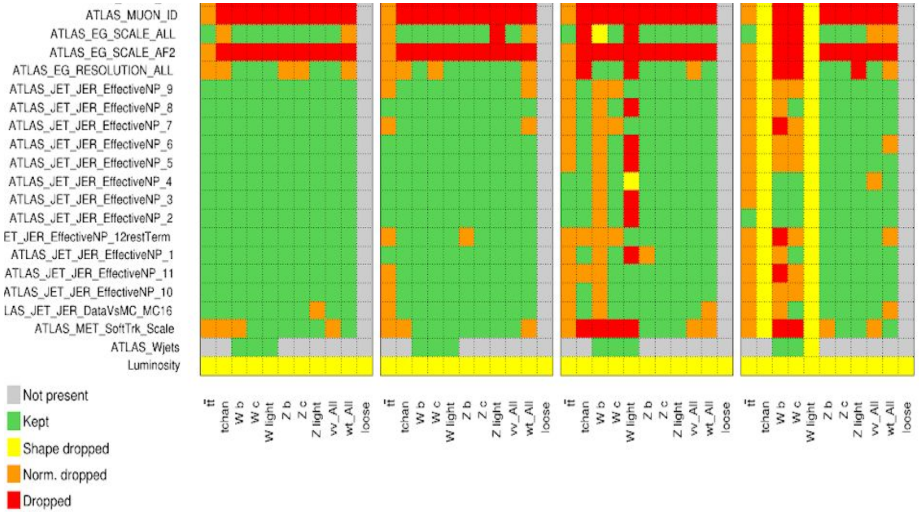


IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

Workspace production and pruning

- Pruning at this step removes negligible effects from systematic variations
  - Control the threshold via
    - SystPruningNorm
    - SystPruningShape
- Study the effect of the pruning you apply!

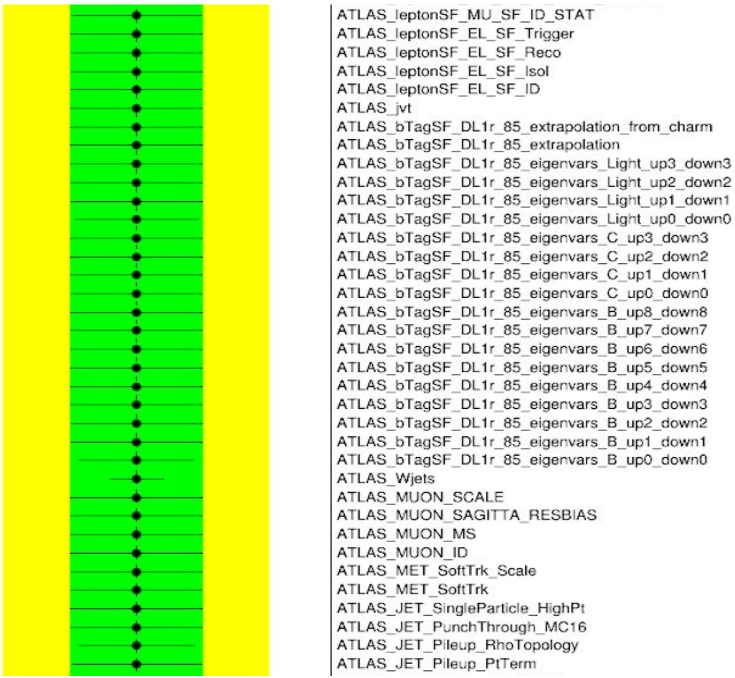


- Not present
- Kept
- Shape dropped
- Norm. dropped
- Dropped

IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

NP pulls, constraints

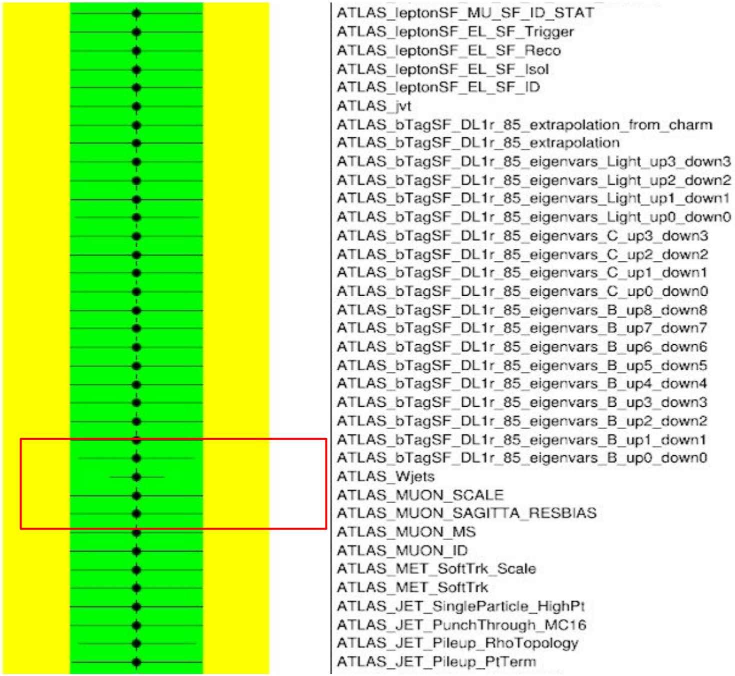


- Useful to monitor NP pulls and constraints:

IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

NP pulls, constraints

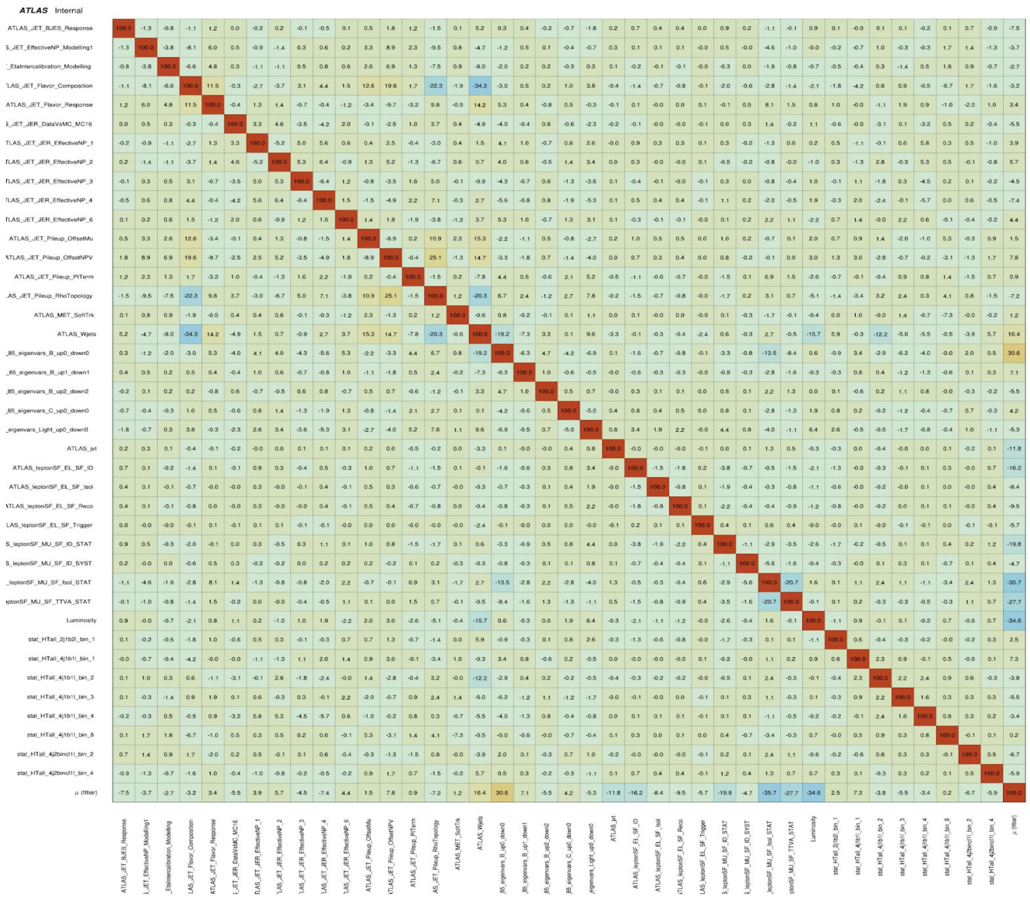


- Useful to monitor NP pulls and constraints
- They are "nuisance", but they can be important!

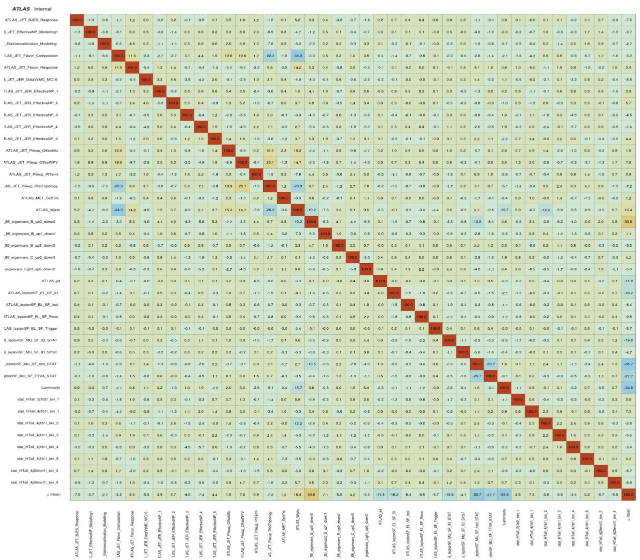
IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

Correlations



Correlations



- 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!

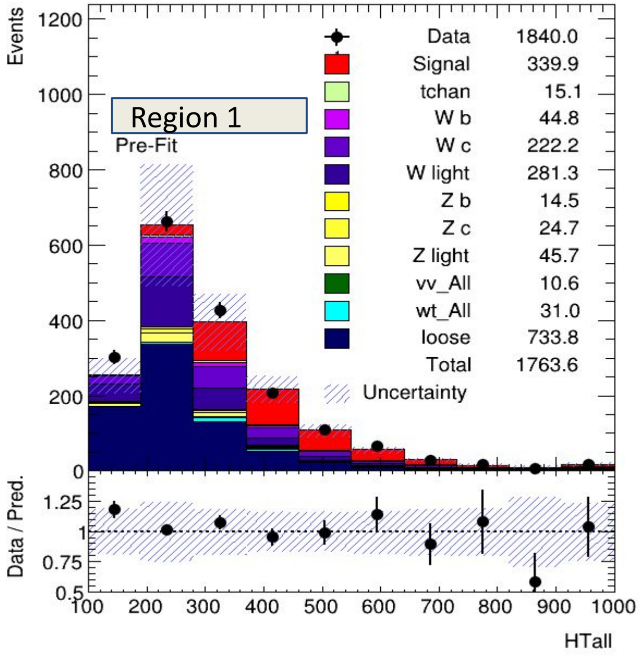
IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ

Inference and post-fit plots

Core framework task: run a profile likelihood fit. Many configuration possibilities:

- Data or Asimov (pseudo-) data
- Including a signal or background-only
- Which regions to include



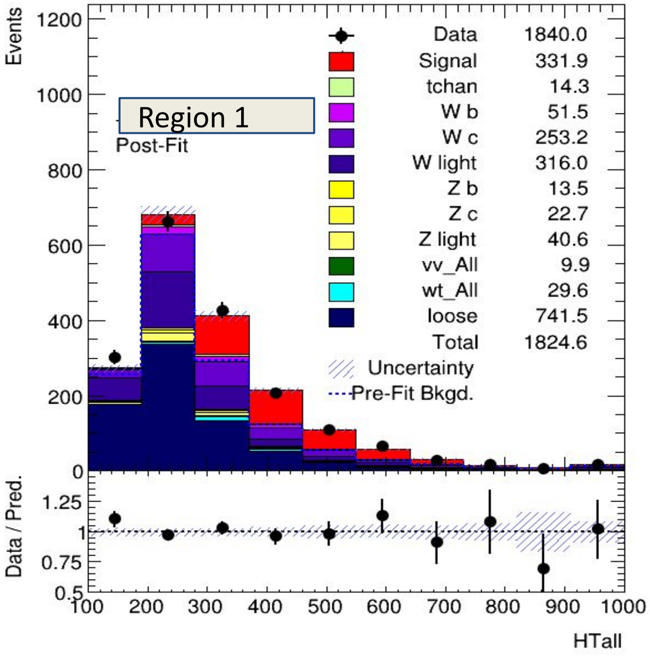
Post-fit plots inference

Core framework task: run a profile likelihood **fit**. Many configuration possibilities:

- Data or Asimov (pseudo-) data
- Including a signal or background-only
- Which regions to include

Many plots and files generated to document and understand the fit

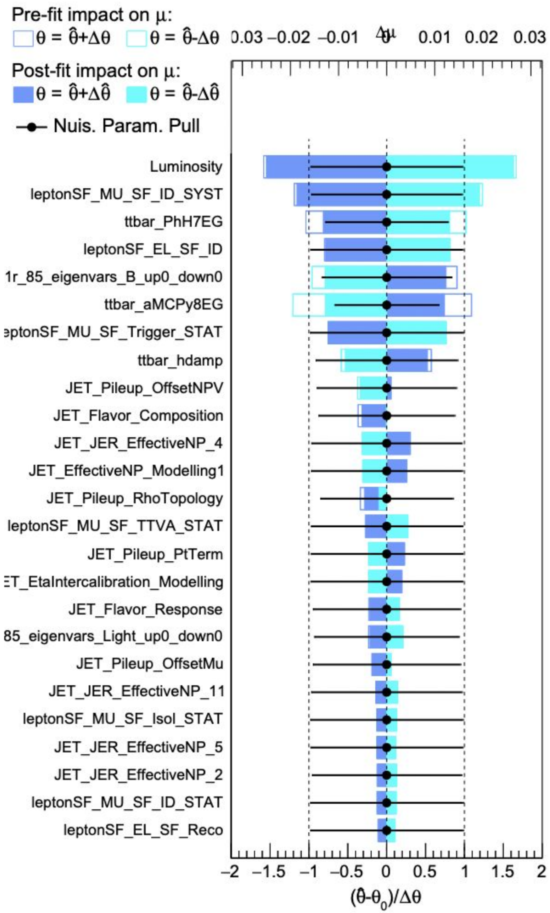
- Best-fit values of all nuisance parameters and associated uncertainties
- Correlations of fit parameters



Impact of NP on the POI (Parameter of Interest)

- To see which nuisance parameter has the largest impact on the uncertainty of our signal strength, we make use of the r action.

- "**Ranking plot**" shows pre-fit and post-fit impact of individual NP on the determination of μ (Parameter of Interest/POI).



Grouped impact table

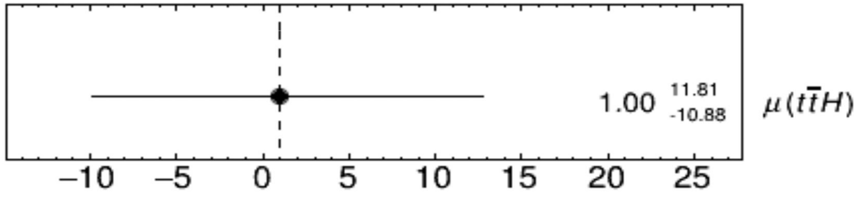
- Besides the ranking feature, **TRExFitter** includes another way of calculating how much certain nuisance parameters "matter".

- The feature discussed here is also called "**grouped impact**". It is particularly useful to evaluate the uncertainty on a parameter of interest (POI) due to a group of nuisance parameters (NPs).

Uncertainty Source	$\Delta\mu$	up	down
EGamma	0.013	0.014	-0.013
FTAG	0.013	0.013	-0.013
JET	0.011	0.011	-0.011
Luminosity	0.026	0.027	-0.025
MET	0.001	0.001	-0.001
Muon	0.023	0.024	-0.022
Modelling	0.024	0.024	-0.023
FullSyst	0.050	0.052	-0.049
Gammas (sim. stat. unc.)	0.009	0.010	-0.009

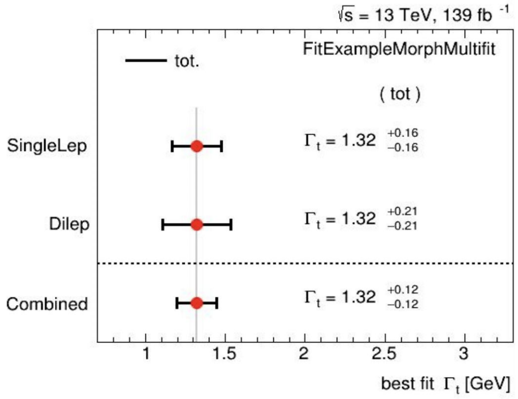
POI (Parameter of Interest)

- The following plot shows you the result for the fit we performed, including the best-fit value for the parameter we ultimately want to extract:



Additional features

- Lots of features implemented beyond a simple fit:
- Combined impact of nuisance parameter groups
- Combination and comparison of different fits
- Toys to evaluate effect of statistical fluctuations in templates defining systematic uncertainties
- Template fitting / morphing
- Exclude nuisance parameters or fix them to specific values
- Correlate or de-correlate nuisance parameters
- Create custom Asimov datasets and fit them



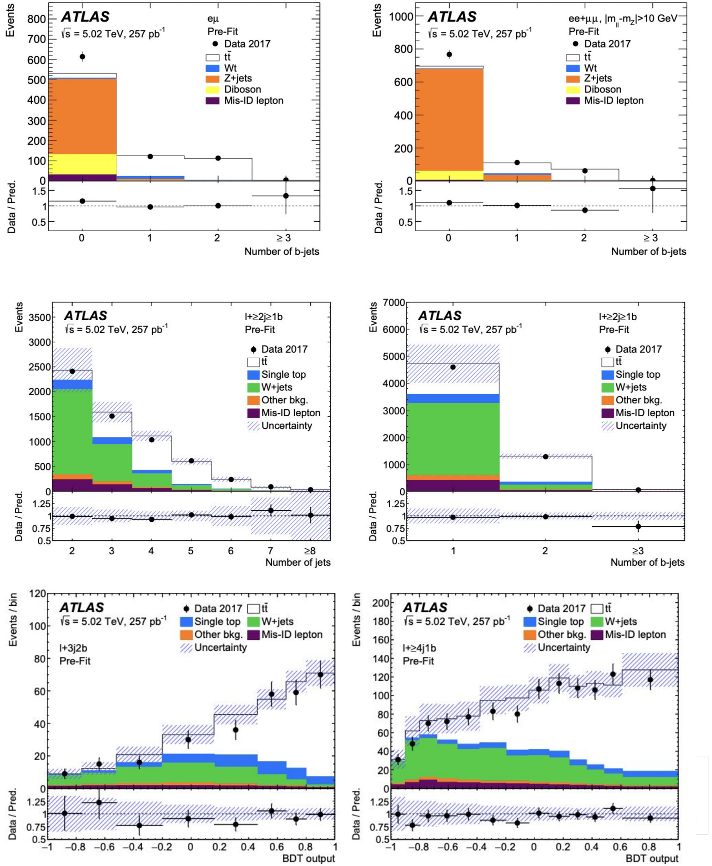
IEAP CTU in Prague SLEZSKÁ UNIVERZITA V OPAVĚ

Publications and TRExFitter

Measurement of the $t\bar{t}$ production cross-section in pp collisions at $\sqrt{s}$ = 5.02 TeV with the ATLAS detector

https://arxiv.org/pdf/2207.01354.pdf

Category	$\delta\sigma_{t\bar{t}}$ [%] Dilepton	Single lepton	Combination
$t\bar{t}$ generator†	1.2	1.0	0.8
$t\bar{t}$ parton-shower/hadronisation*,†	0.3	0.9	0.7
$t\bar{t}$ h_{damp} and scale variations†	1.0	1.1	0.8
$t\bar{t}$ parton distribution functions†	0.2	0.2	0.2
Single-top background	1.1	0.8	0.6
W/Z + jets background*	0.8	2.4	1.8
Diboson background	0.3	0.1	< 0.1
Misidentified leptons*	0.7	0.3	0.3
Electron identification/isolation	0.8	1.2	0.8
Electron energy scale/resolution	0.1	0.1	< 0.1
Muon identification/isolation	0.6	0.2	0.3
Muon momentum scale/resolution	0.1	0.1	0.1
Lepton-trigger efficiency	0.2	0.9	0.7
Jet-energy scale/resolution	0.1	1.1	0.8
$\sqrt{s}$ = 5.02 TeV JES correction	0.1	0.6	0.5
Jet-vertex tagging	< 0.1	0.2	0.2
Flavour tagging	0.1	1.1	0.8
$E_{\text{T}}^{\text{miss}}$	0.1	0.4	0.3
Simulation statistical uncertainty*	0.2	0.6	0.5
Data statistical uncertainty*	6.8	1.3	1.3
Total systematic uncertainty	3.1	4.2	3.7
Integrated luminosity	1.8	1.6	1.6
Beam energy	0.3	0.3	0.3
Total uncertainty	7.5	4.5	3.9



Summary

- TRExFitter, a very powerful and configurable tool
- User friendly and not a black-box
- Used in many physics analysis including our recent ongoing top-quark pair production in proton-lead collisions

Thank You

HI top analysis meeting

Test statistics

- **Best-fit result** (unconditional maximum likelihood estimate) of measurement
  - Maximize likelihood by varying $\vec{k}$, $\vec{\theta}$, POI μ is part of $\vec{k}$

- For significance / limit, make use of **profile likelihood ratio** (reference: arXiv:1007.1727)
  - In asymptotic limit (more than ~10 events / bin), can quickly calculate significance/limits

maximum likelihood for given μ

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

unconditional maximum likelihood

- **Discovery significance**: $q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 , \\ 0 & \hat{\mu} < 0 , \end{cases}$
  - $Z_0 = \sqrt{q_0}$
  - Takes two fits to calculate

- **Upper parameter limits**: $\tilde{q}_\mu$ test statistic, $\tilde{q}_\mu = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0 , \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & 0 \leq \hat{\mu} \leq \mu , \\ 0 & \hat{\mu} > \mu . \end{cases}$
  - We also use the CL_S method (reference)
  - Vary μ to find the CL_S = 5% crossing for 95% parameter limits

IEAP CTU in Prague

SLEZSKÁ UNIVERZITA V OPAVĚ