RooStatsCms: a tool for analyses, modelling, combination and statistical studies

D. Piparo, G. Schott, G. Quast

Institut für Experimentelle Kernphysik
Universität Karlsruhe
Outline

• The need for a tool for statistical methods and channels combination

• A possible solution: RooStatsCms

• Benchmark analysis: H→ττ (VBF)

• The “modified frequentist” method
  – Significance
  – SM cross-section exclusion

• The “profile likelihood” method
  – Upper limits
The need for a tool

- Reliable implementation of multiple statistical methods

- Combine analyses:
  - Information lies more at the analysis level than at the result level
  - Consistent treatment of constraints and their correlations: no double counting
  - Stronger limits on quantities like Higgs production cross section, mass ...

- Do not replace existing analyses but complement their results

  **Crucial especially in the early phases of the data taking**

- Easy user interface

- Satisfactory documentation
A possible solution: RooStatsCms (RSC).

Based on RooFit:
- Originally developed in BaBar, used in many experiments/collaborations
- Part of standard ROOT distribution

RSC runs on a laptop.

Three parts:
- Modelling and combination
- Statistical methods (based on likelihood ratios)
- Advanced graphic routines

Doxy documentation of every class, method and member.

It comes with CINT dictionaries (macros, interactive root).

Available to CMS at: [www-ekp.physik.uni-karlsruhe.de/~RooStatsCms](http://www-ekp.physik.uni-karlsruhe.de/~RooStatsCms).
- Visit [tinyurl.com/rscpasswd](http://tinyurl.com/rscpasswd) for username and password
- More material in the CMS Wiki
- Statistical methods and graphic routines public: [www-ekp.physik.uni-karlsruhe.de/~RooStatsKarlsruhe](http://www-ekp.physik.uni-karlsruhe.de/~RooStatsKarlsruhe)
- RooStatsKarlsruhe: part of the negotiations towards a common tool with Atlas

RSC: in “production phase”
- Workshop at CERN in June
- Approved results
• Build a complete combined analysis model from ASCII datacards ("config files")
  – Background and signal components of each analysis
  – Shapes from parametrisation or histos
  – Constraints and their correlations
  – Basic syntax: include, if ...
  – Two lines of C++ to produce the RooFit Pdf

• Datacard advantages:
  – Automatic bookkeeping of what is done
  – Factorise model from C++ code
  – Easy to share

ASCII Card
2 analyses

RscCombinedModel mymodel ("hzz4l");
RooAbsPdf* sb_pdf=mymodel.getPdf();
• Yields can be expressed as products of different terms:
  – Branching Ratios
  – Efficiencies
  – Cross section
  – Luminosity
  – $\sigma_H/\sigma_{SM}$

Yield = $BR \cdot \varepsilon \cdot \sigma_{\text{Prod}} \cdot \text{Lumi} \cdot \sigma_H/\sigma_{SM}$

• Each term: systematics can be included
• The same applies also to shape parameters
• Relate terms from one analysis to the other with correlations
Combination example: PTDR 30 fb$^{-1}$

- Reproduced analysis of PTDR: $H \to ZZ \to 4l$ and $H \to \gamma\gamma$
  - (bkgs $H \to ZZ$ 100% correlated)
- Added combination of $H \to ZZ \to 4l$ and $H \to \gamma\gamma$
  - counting and non counting experiment: symmetrical treatment
- Significance estimator: $\sqrt{2\ln Q}$
- Variable $Q = L_{s+b}/L_b$ with $L_{s+b}, L_b$ likelihoods in the sig+bkg and bkg only hypotheses
• Perform a statistical analysis of your result

• RSC statistical methods: based on likelihood ratios

• Two statistical methods well tested:
  – The -2lnQ distributions for hypothesis separation
  – The Profile Likelihood method

• Sometimes analysis time-consuming (lots of toy-MC experiments):
  – “Batch friendly”: sum up your results

• Easy to get out of results plots in a presentation-ready form
• Organisation of the classes:

Statistical Methods – Mother: StatisticalMethod
- LimitCalculator
- PLScan
- FCCalculator

Statistical Methods Results – Mother: StatisticalResult
- LimitResults
- PLScanResults
- FCResults

Statistical Plot – Mother: StatisticalPlot
- LimitPlot
- PLScanPlot (add also FC curves)

Constraints
Mother: NLLPenalty.cc
- Constraint.cc
- ConstrBlock2.cc
- ConstrBlock3.cc
- ConstrBlockArray.cc

+ LEPBandPlot
+ ExclusionBandPlot
Marginalisation

- MC phase-space integration
- Lots of toy experiments

Profiling

- Penalty term in the likelihood (\( \log L_{T} = \log L_{\text{base}} + \log L_{\text{Penalty}} \)) e.g.

  1. One uncorrelated Gaussian constraint
  \[
  \log L_{P} \sim 0.5 \cdot \frac{(m-m_{0})^{2}}{\sigma_{m}^{2}}
  \]

  2. Correlated Gaussian constraints
  \[
  \log L_{P} \sim 0.5 \cdot (m-m_{0})^{\top} \cdot V^{-1} \cdot (m-m_{0}), \text{ V is correlation matrix}
  \]

- No toys: go for a few fits

High statistics/ Gaussian case: two methods converge
Benchmark analysis: VBF $H \rightarrow \tau\tau$

- Used as benchmark for the tool
- Results approved by the CMS collaboration
- Vector boson fusion
- Integrated lumi: 1 fb$^{-1}$
- Small signal on a significant background
- No discovery expected with this lumi
- Four mass hypotheses:
  - 115, 125, 135, 145 GeV

<table>
<thead>
<tr>
<th>Mass</th>
<th>N Sig</th>
<th>N Bkg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(12% sys)</td>
<td>(30% sys)</td>
</tr>
<tr>
<td>115</td>
<td>1.6</td>
<td>45.2</td>
</tr>
<tr>
<td>125</td>
<td>1.4</td>
<td>45.2</td>
</tr>
<tr>
<td>135</td>
<td>1.1</td>
<td>45.2</td>
</tr>
<tr>
<td>145</td>
<td>0.6</td>
<td>45.2</td>
</tr>
</tbody>
</table>
Separation of s+b and b only

- Idea: separation of hypotheses using the likelihoods ratio, $Q$, assuming signal+background (“s+b”) and the background-only “b” hypotheses, as test statistic.

- Consider “P-values” (also called 1-CL$_{S+B}$, 1-CL$_B$) of -2lnQ distributions obtained from s+b and b samples.

**Treatment of systematics:**

For every toy MC experiment, before the generation of the toy dataset, parameters affected by systematics are properly fluctuated.
• $\text{CL}_B$: background CL, measure of the compatibility of the experiment with the B-only hypothesis

• $1 - \text{CL}_B$: probability for a B-only experiment to give a more S+B-like likelihood ratio than the observed one

• Correspondence between $1 - \text{CL}_B$ and the resulting significance (Gaussian approximation):
  - # of standard deviations of an (assumed) Gaussian distribution of the background.
  - Take $\text{CL}_B$ assuming the expected s+b yield (i.e. median $-2\ln Q$ for s+b distribution)

• $\text{CL}_{S+B}$: measure of the compatibility of the experiment with the S+B hypothesis
  if CL is small ( < 5% ) the S+B hypothesis can be excluded at more than 95% CL but it does not mean that the signal hypothesis is excluded at that level

**Modified frequentist approach:**
take $\text{CL}_S$ the signal significance, to be: $\text{CL}_S \equiv \text{CL}_{S+B} / \text{CL}_B$ (heavily used by LEP, HERA and TEVATRON experiments)
Significance calculated for the $H \rightarrow \tau \tau$ analysis using $1-\text{CL}_B$

In this case significance does not tell us much.

The question becomes:

"Which production cross section can I exclude with the data I have?"
Assume to observe the expected background (i.e. median of the background distribution) and no signal

- Amplify the SM production cross section by a factor necessary to obtain $\text{CL}_s=0.05$ → “95% exclusion”

$\sim 80$ h on one CPU

**Bands:**
- Assume to observe $N_b + n \cdot \sqrt{N_b}$, where $n=2,1,-1,-2$ for the -2,-1,1,2 sigma band border respectively
- Systematics taken into account in distributions of $-2\ln Q$ (marginalisation)
The “profile likelihood” method

- Likelihood scanned w.r.t. a variable (e.g. signal yield)
- At each point, partial likelihood maximized w.r.t. nuisance parameters
- Intersection with horizontal lines gives upper limits / two sided intervals
- Systematics taken into account with penalty terms in the Likelihoods (profiling)

See PLCalculator, PLResults, PLPlot documentation
• Again VBF $H \rightarrow \tau\tau$ as benchmark (no systematics here)
• With profile likelihood the 95% CL UL is 10.71 events = 6.7 SM cross section
  – to compare to ~5.5 with CL$_s$
• Coverage: frequency in which in of toy experiments the “real” value is included in the confidence interval
• Coverage tested with several MC toys experiments:
  – For low signal yields, the profile likelihood method largely over-covers
  – The method works well for large signal (and luminosity)
Conclusions

- RooStatsCms - tool for statistical studies and analyses combination in the CMS collaboration
- Implemented and tested existing and widely accepted statistical methods: in 'production phase'
- Study of VBF $H \rightarrow \tau \tau$ carried out:
  - SM production cross section exclusion power
  - PL likelihood upper limits
- The tool has been 'adopted' by the Higgs WG; it will be used for the Higgs results.
- Extensive X-checks done or planned
- It became solid tool
  - Example macros, documentation, CMS workshop, …
- Integration in ROOT being discussed
- Working on documenting the tool and used methods in a support document

Future Plans

- Continue crosschecking with independent tools
- Add other statistical methods (Working on a full frequentist approach)
- Improve MC integration technique and numerical procedures
  - Such as approach based on Markov chains Monte Carlo