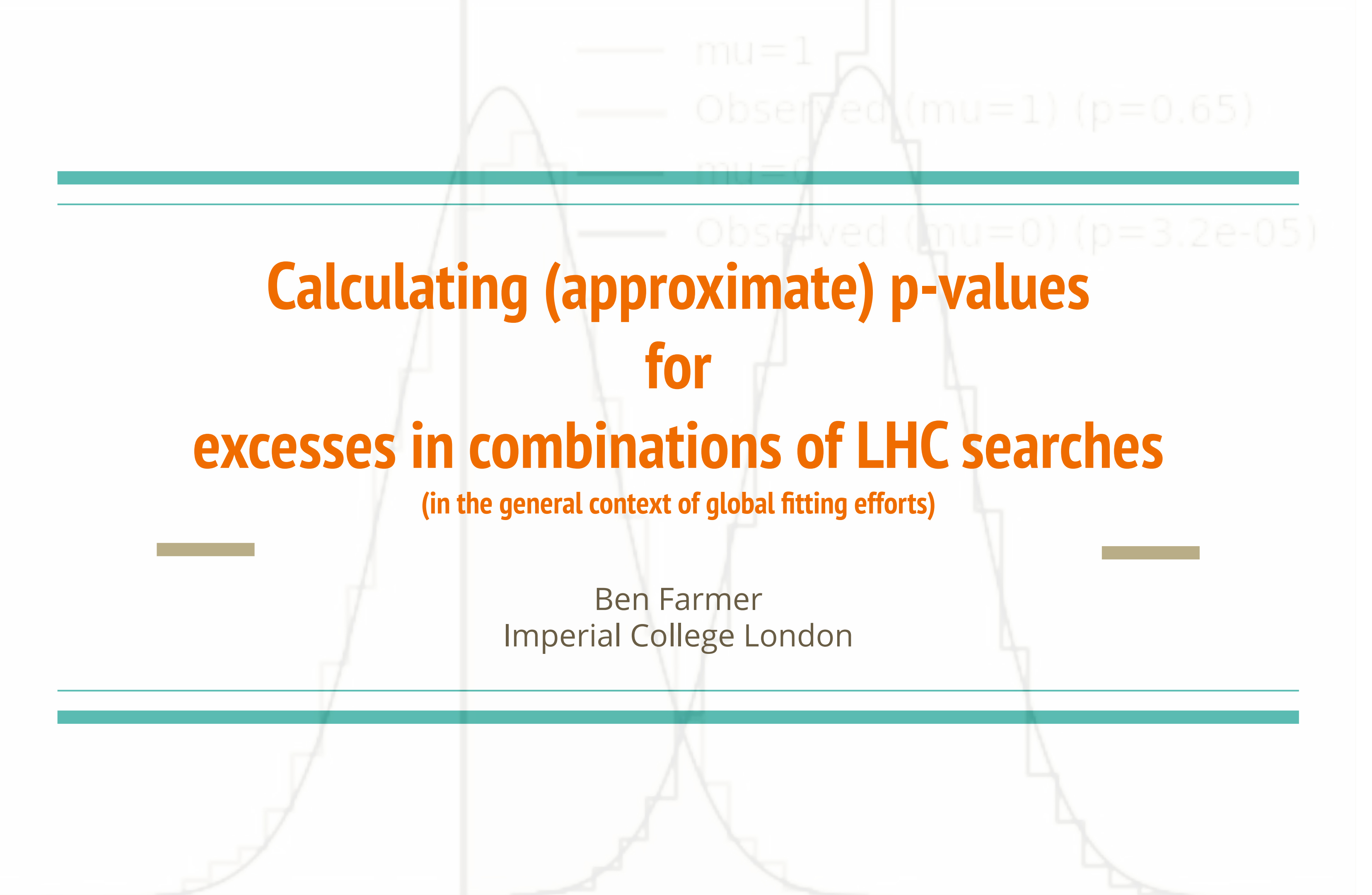


Calculating (approximate) p-values for excesses in combinations of LHC searches

(in the general context of global fitting efforts)

Ben Farmer
Imperial College London

Outline

- Why calculate a dedicated “discovery” p-value? (Some problems with asymptotic confidence regions, and the “look-elsewhere” effect)
- “Local” p-values
- “Too global” p-values
- Example
- Summary

Statistics goals of LHC reinterpretation

- Rule out models

Statistics goals of LHC reinterpretation

- ~~Rule out models~~
- Discover new physics!

Statistics goals of LHC reinterpretation

- ~~Rule out models~~
- Discover new physics!

How? Ideally, rule out Standard Model at “five sigma” level

But not all statistical tests are created equal!

Different tests use different test statistics, so evaluate different data as statistically significant or not!

Notion of “more extreme than observed” depends on how we choose to order the possible data outcomes!

Trivial example: coin flips

Consider the following sequence of flips:

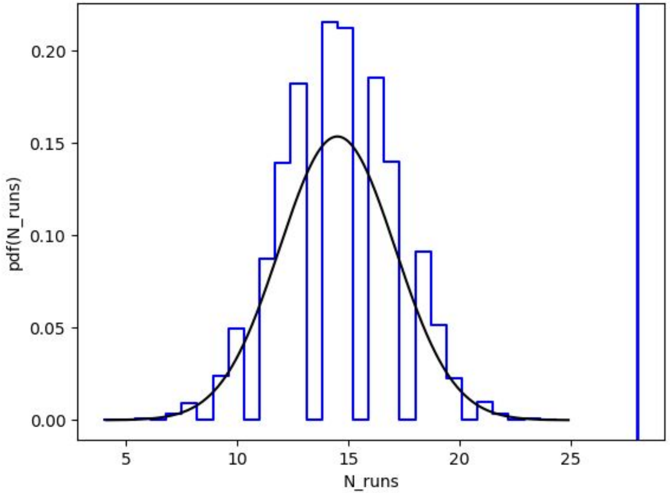
HTHTHTHTHTHTHTHTHTHTHTHTHTHT

N:28 (H: 14 , T: 14)

Pr(H>=14|N=28) ~ 0.6 ~ 0σ

R = #runs of same outcome = 28

Pr(R>=28|N=28) = 1e-07 ~ 5.2σ



“Standard practice” …at least in global fitting

$$\Lambda = -2\log\left(\frac{L(\mathrm{data}|\theta,\hat{\hat{\eta}})}{L(\mathrm{data}|\hat{\theta},\hat{\eta})}\right)$$

Go to theory, with parameters ϴ

Pick 2D parameter plane(s) of interest

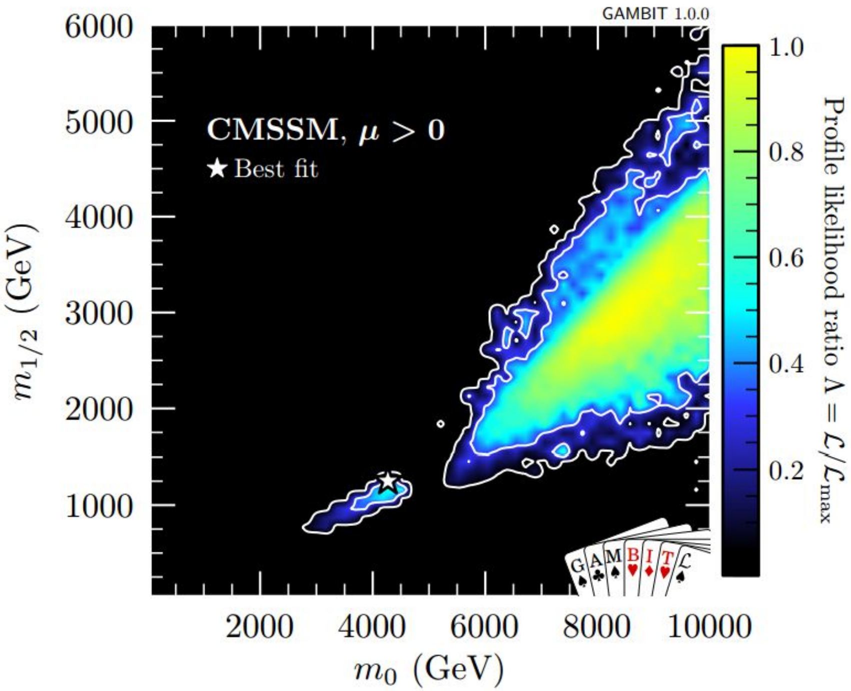
Construct confidence regions via likelihood ratio test

Wilks’ theorem -> asymptotically distributed as chi-squared with 2 DOF

Draw contour around region where $\Lambda < 5.99$

Outside this is “excluded at 95% confidence”

If Standard Model is nested in ϴ, exclude if outside X sigma interval?



arXiv:1705.07935: *Global fits of GUT-scale SUSY models with GAMBIT*

Problems for discovery

This procedure is simple and computationally cheap to implement.

But:

- Reliance on asymptotic theory (often fails in theory parameter spaces)
- “Look elsewhere effect” (construct many confidence regions in many parameter projections)

(We tend to accept this construction more by convention than because we really believe the coverage is correct! That’s ok, but if we want to claim that a signal might exist then we need a higher standard)

Asymptotic theory

~ requires MLE to be normally distributed

Wilks' theorem fails if:

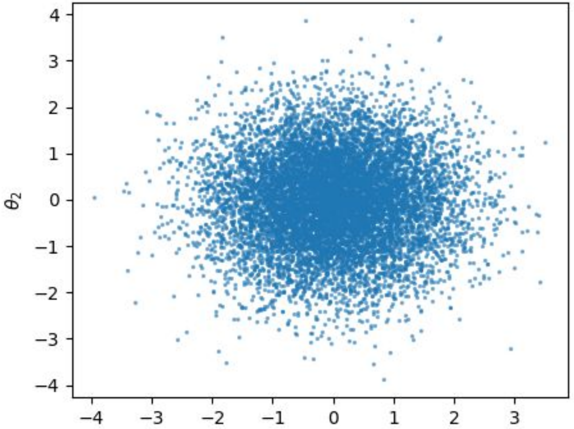
- MLE too close to boundary
- Parameter space not "regular" (basically, each possible parameter choice must uniquely identify a data pdf. No degeneracies allowed!)
- Likelihood surface not sufficiently differentiable
- Support of data depends on parameters

Asymptotic theory

~ requires MLE to be normally distributed

Wilks’ theorem fails if:

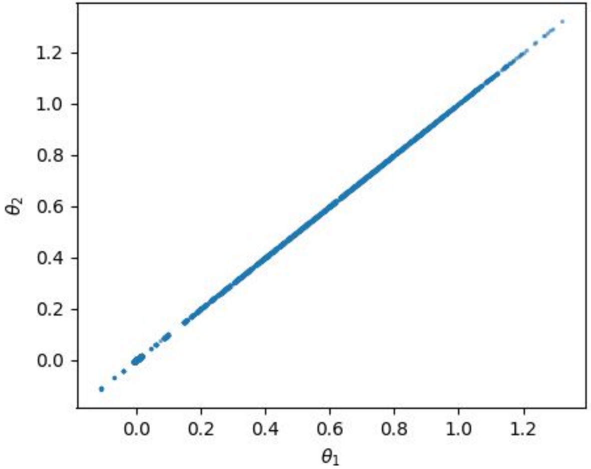
- MLE too close to boundary
- Parameter space not “regular” (basically, each possible parameter choice must uniquely identify a data pdf. No degeneracies allowed!)
- ~~Likelihood surface not sufficiently differentiable~~
- ~~Support of data depends on parameters~~

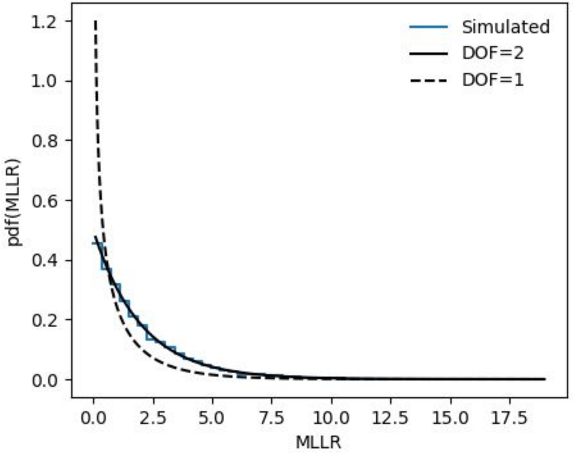


Data:
2 normal random variables

$$\mu_1 = \theta_1$$

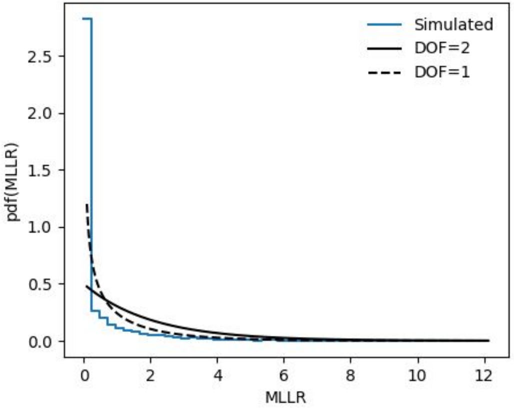
$$\mu_2 = \theta_2$$

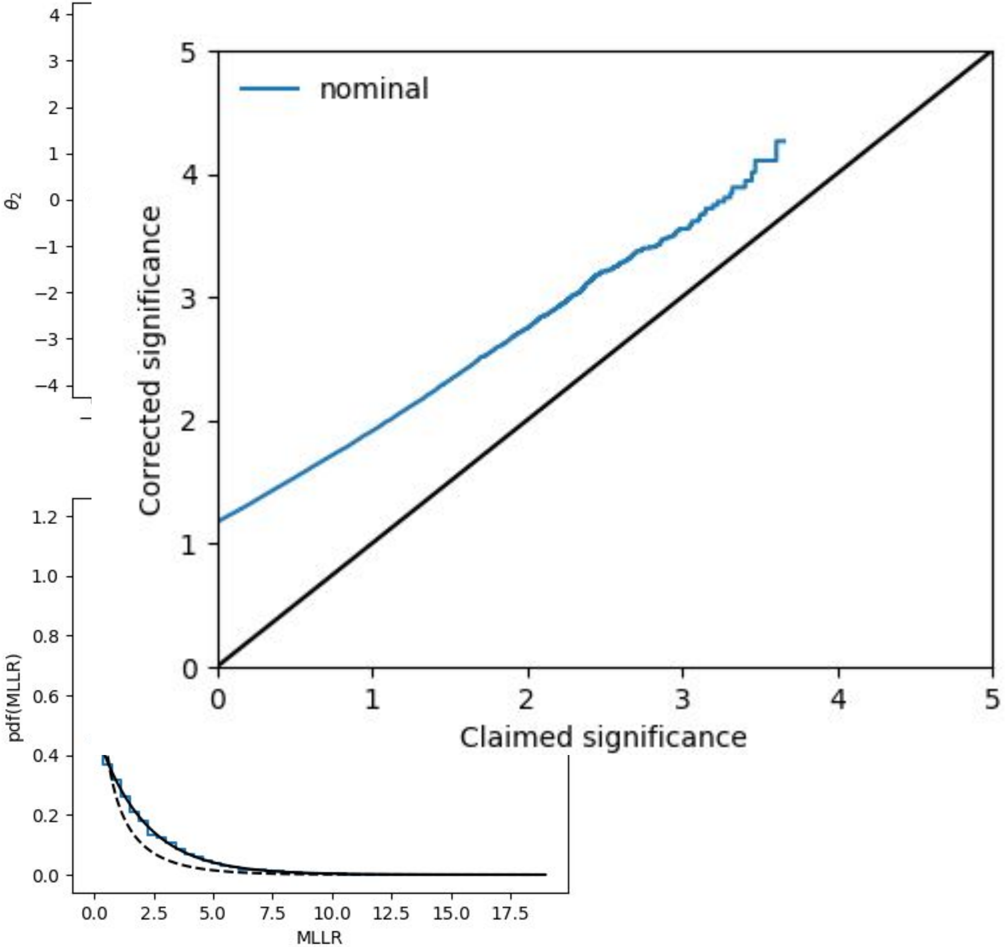




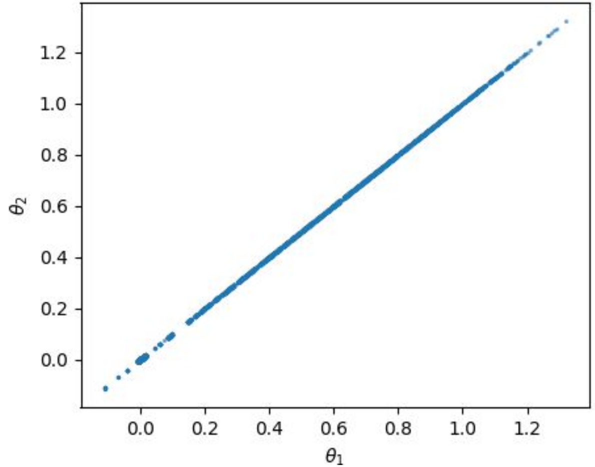
Data:
1 normal random variables

$$\mu_1 = \theta_1^2 + \theta_2^2$$



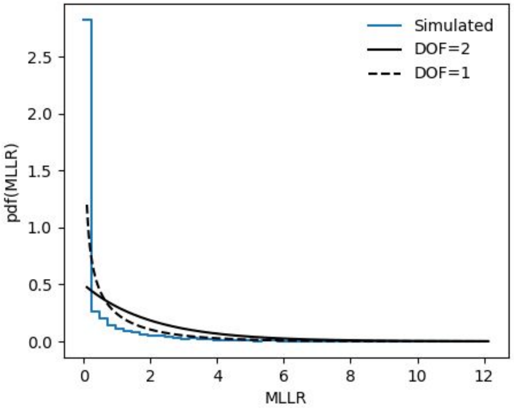

nominal
Corrected significance
Claimed significance
0
1
2
3
4
5
θ_2
4
3
2
1
0
−1
−2
−3
−4
pdf(MLLR)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
MLLR

s


θ_2
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
θ_1

es

θ_2^2


Simulated
DOF=2
DOF=1
pdf(MLLR)
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
MLLR

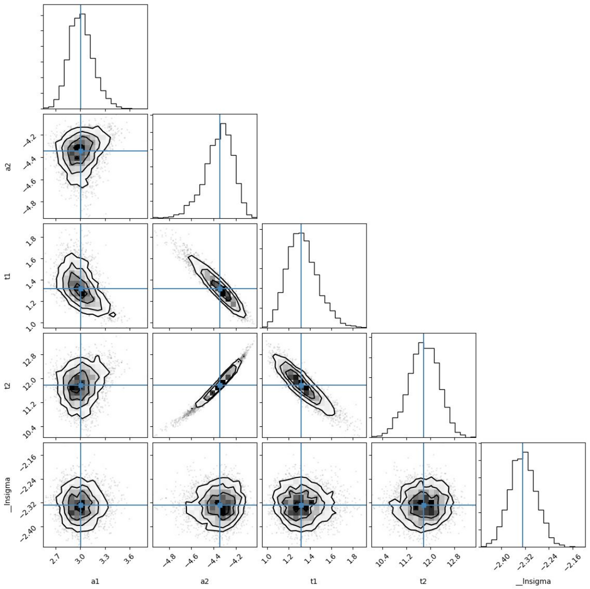
“Look elsewhere effect”

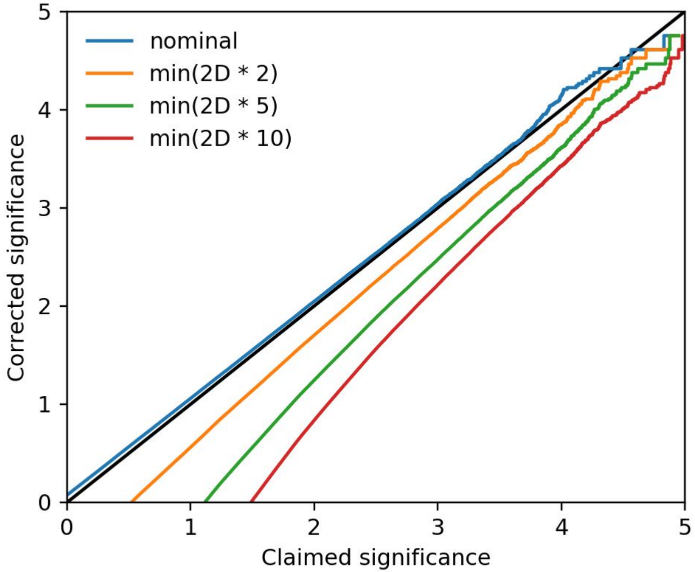
Suppose we construct many confidence regions for various projections of a large parameter space…

…but then see a “2 sigma” deviation in one of them

-> CRs don’t know about each other! Effectively we are doing:

$$N = \max(N_1, N_2, ..., N_n)$$





arxiv: 1806.02293 - *Search for chargino–neutralino production using recursive jigsaw reconstruction in final states with two or three charged leptons in proton–proton collisions at √ s = 13 TeV with the ATLAS detector*

LHC analyses

At the most basic: Poisson counting experiments

Have some number of “signal regions” (defined by event selection criteria)

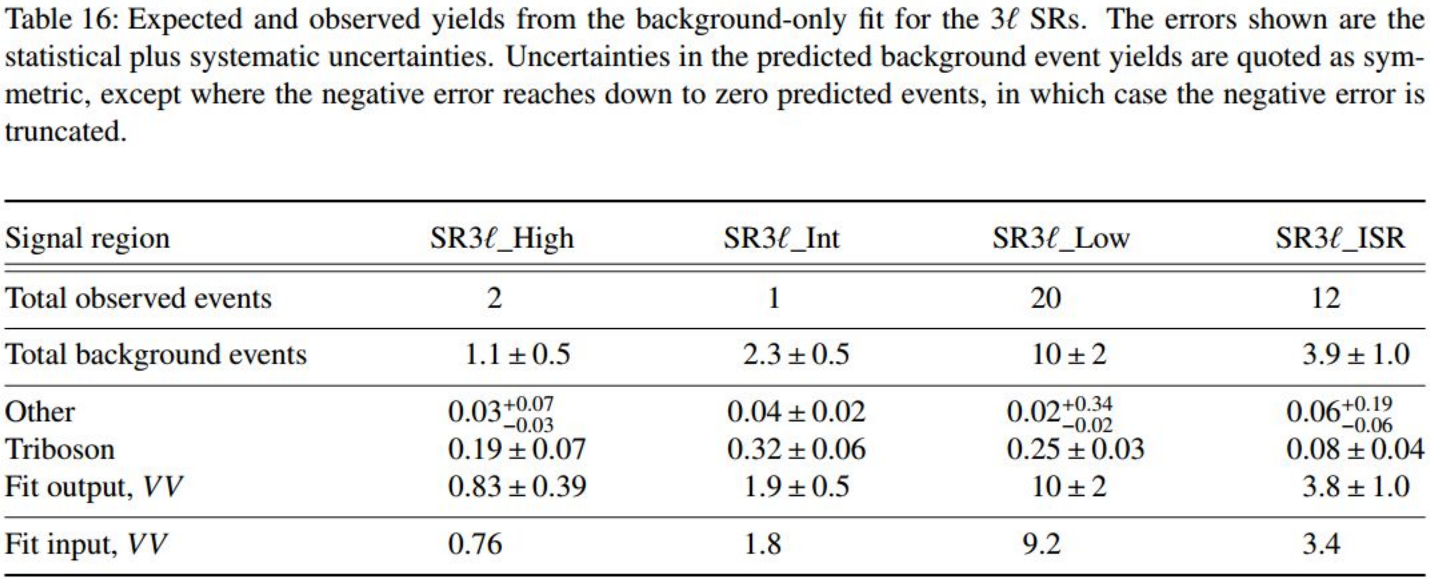


Table 16: Expected and observed yields from the background-only fit for the 3ℓ SRs. The errors shown are the statistical plus systematic uncertainties. Uncertainties in the predicted background event yields are quoted as symmetric, except where the negative error reaches down to zero predicted events, in which case the negative error is truncated.

Signal region	SR3ℓ_High	SR3ℓ_Int	SR3ℓ_Low	SR3ℓ_ISR
Total observed events	2	1	20	12
Total background events	1.1 ± 0.5	2.3 ± 0.5	10 ± 2	3.9 ± 1.0
Other	$0.03^{+0.07}_{-0.03}$	0.04 ± 0.02	$0.02^{+0.34}_{-0.02}$	$0.06^{+0.19}_{-0.06}$
Triboson	0.19 ± 0.07	0.32 ± 0.06	0.25 ± 0.03	0.08 ± 0.04
Fit output, VV	0.83 ± 0.39	1.9 ± 0.5	10 ± 2	3.8 ± 1.0
Fit input, VV	0.76	1.8	9.2	3.4

LHC analyses

P-values?

$$n = 10$$

$$b = 3.9$$

$$p = \sum_{k=n}^{\infty} \frac{b^k e^{-b}}{k!} \approx 0.0002$$

$$\rightarrow 3.5\sigma$$

(ignoring background uncertainty)

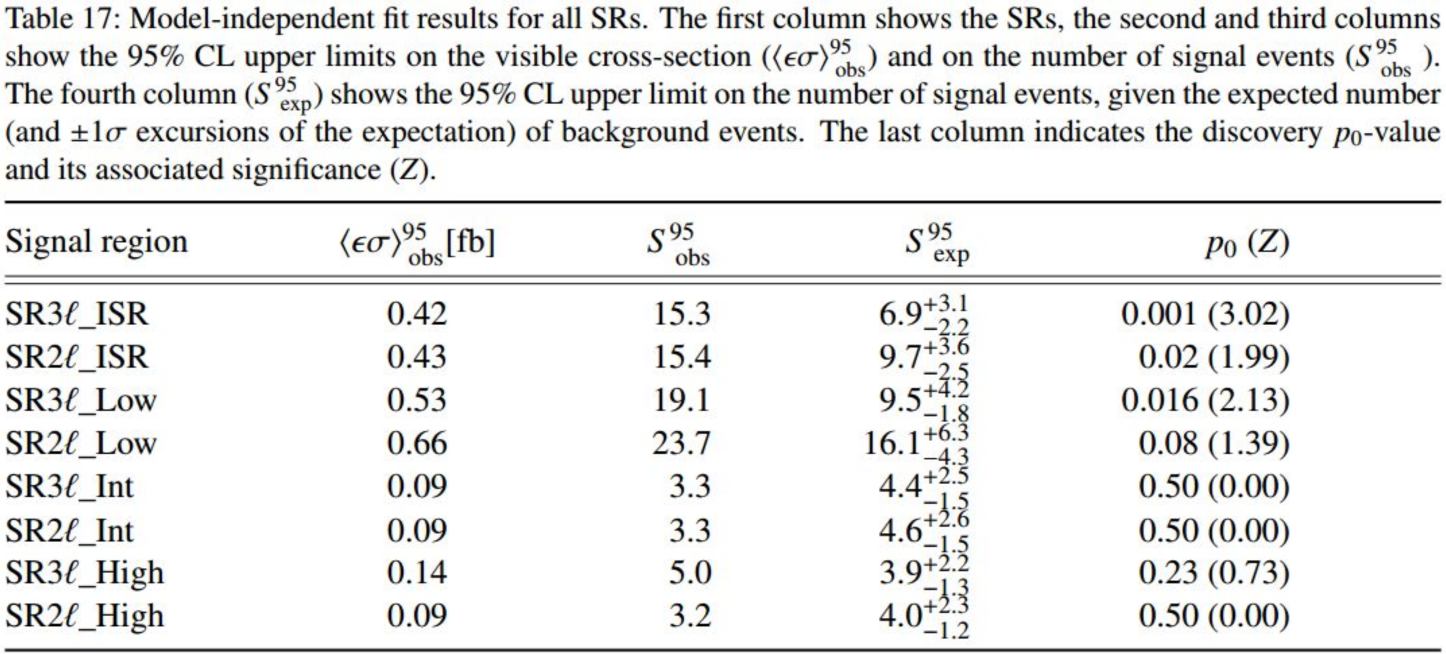
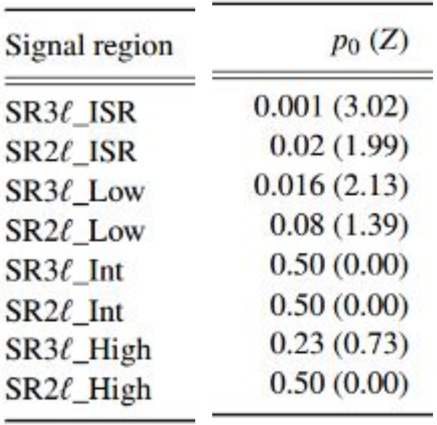


Table 17: Model-independent fit results for all SRs. The first column shows the SRs, the second and third columns show the 95% CL upper limits on the visible cross-section ($\langle\epsilon\sigma\rangle^{95}_{\mathrm{obs}}$) and on the number of signal events (S^{95}_{obs}). The fourth column (S^{95}_{exp}) shows the 95% CL upper limit on the number of signal events, given the expected number (and $\pm 1\sigma$ excursions of the expectation) of background events. The last column indicates the discovery p_0-value and its associated significance (Z).

Signal region	$\langle\epsilon\sigma\rangle^{95}_{\mathrm{obs}}$[fb]	S^{95}_{obs}	S^{95}_{exp}	p_0 (Z)
SR3ℓ_ISR	0.42	15.3	$6.9^{+3.1}_{-2.2}$	0.001 (3.02)
SR2ℓ_ISR	0.43	15.4	$9.7^{+3.6}_{-2.5}$	0.02 (1.99)
SR3ℓ_Low	0.53	19.1	$9.5^{+4.2}_{-1.8}$	0.016 (2.13)
SR2ℓ_Low	0.66	23.7	$16.1^{+6.3}_{-4.3}$	0.08 (1.39)
SR3ℓ_Int	0.09	3.3	$4.4^{+2.5}_{-1.5}$	0.50 (0.00)
SR2ℓ_Int	0.09	3.3	$4.6^{+2.6}_{-1.5}$	0.50 (0.00)
SR3ℓ_High	0.14	5.0	$3.9^{+2.2}_{-1.3}$	0.23 (0.73)
SR2ℓ_High	0.09	3.2	$4.0^{+2.3}_{-1.2}$	0.50 (0.00)

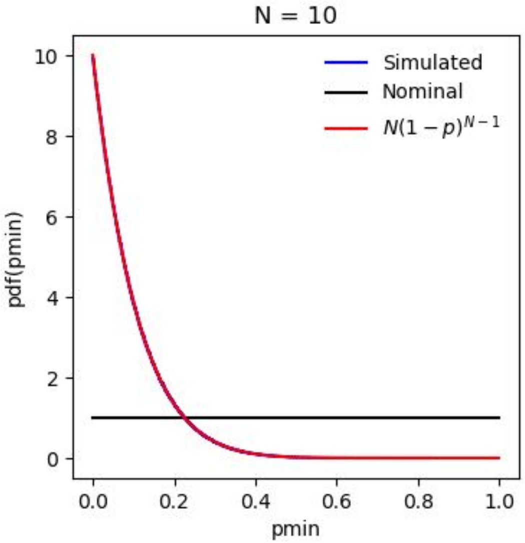
Look-elsewhere effect again

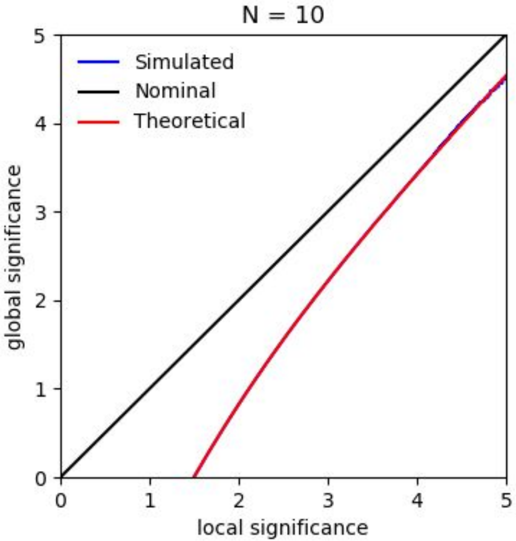


Signal region	p_0 (Z)
SR3ℓ_ISR	0.001 (3.02)
SR2ℓ_ISR	0.02 (1.99)
SR3ℓ_Low	0.016 (2.13)
SR2ℓ_Low	0.08 (1.39)
SR3ℓ_Int	0.50 (0.00)
SR2ℓ_Int	0.50 (0.00)
SR3ℓ_High	0.23 (0.73)
SR2ℓ_High	0.50 (0.00)

…But perhaps not so surprising to see a 3 sigma excess in one signal region, when we have many signal regions!

→ “Trial correction” needed.





“Local” p-values

- Relatively clear how to do this (physics part still tricky though, see e.g. talk by Anders Kvellestad)
- Good asymptotic behaviour (generally)
- Test null hypothesis against one signal hypothesis

Ref: arXiv:1007.1727, G. Cowan et al.

“Asymptotic formulae for likelihood-based tests of new physics”

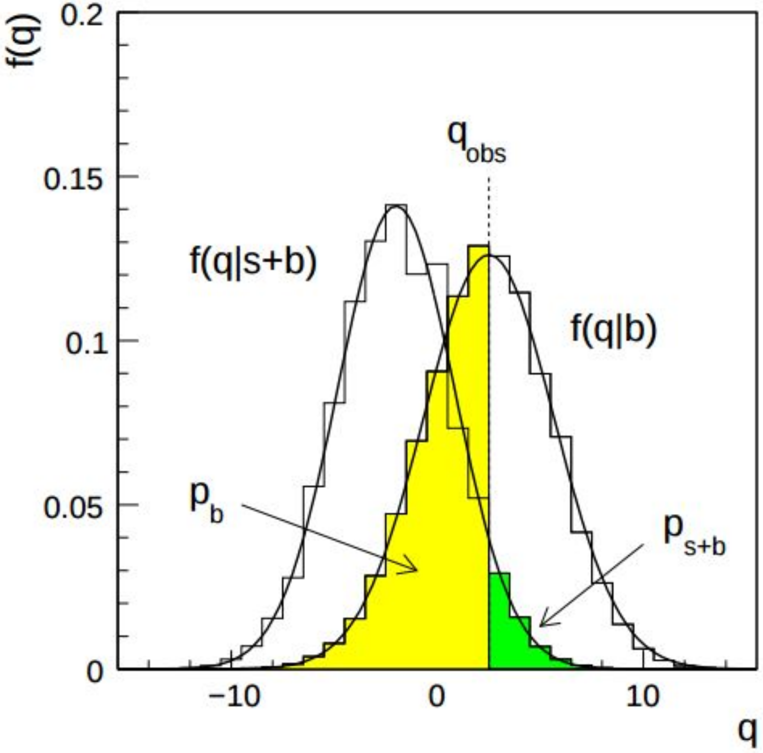
>3000 citations

Example:

$$q = -2 \ln \frac{L_{s+b}}{L_b} \,,$$

$$\lambda_i = \mu s_i(\phi) + b_i + \theta_i$$

$$q = -2 \ln \frac{L(\mu = 1, \hat{\hat{\boldsymbol{\theta}}}(1))}{L(\mu = 0, \hat{\hat{\boldsymbol{\theta}}}(0))}$$



Note: this is where combination comes in! Same whether combining signal regions or whole analyses:

$$L(\mu) = \prod_i L_i(\mu)$$

Figure 6: The distribution of the statistic $q = -2\ln(L_{s+b}/L_b)$ under the hypotheses of $\mu = 0$ and $\mu = 1$ (see text).

*Note: for simple vs simple hypothesis testing, the likelihood ratio gives the best discrimination (power, **Neyman-Pearson Lemma**)

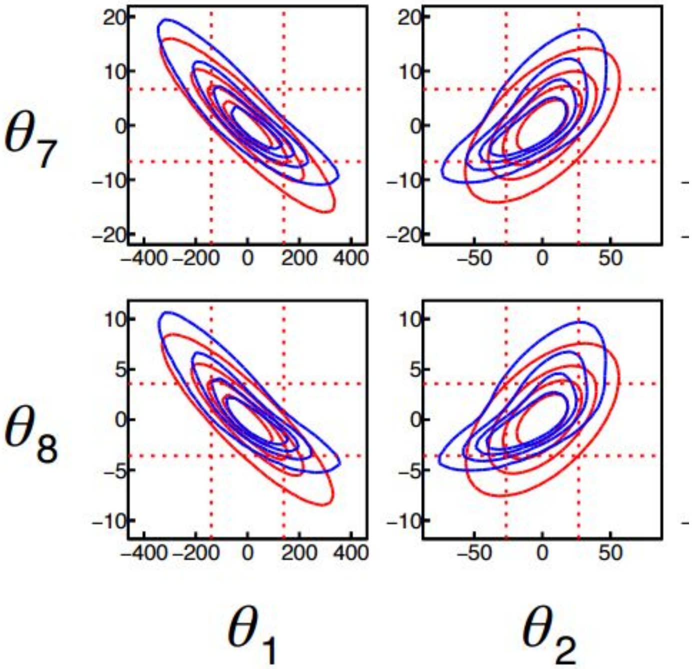
- So far this is all standard to experimentalists
- Critically, it requires a *signal model* of some sort:

$$\lambda_i = \mu s_i(\theta) + b_i$$

  - This is where reinterpretations can have a big role to play!
  - Remember, new signal model = new test statistic
    - Can get high significance where before there was nothing!
      - A prima-facie non-significant set of deviations across many analyses can in fact be highly significant if a model predicts them (and doesn't predict too much else)
- Generically speaking, then, to compute local significances of excesses in a combination of e.g. ATLAS + CMS searches, we **require** BSM theories, and must be able to compute their signal predictions for the relevant searches (see again Anders Kvellestad's talk for how we do this in GAMBIT!).

Experimental input

- As well as signal model calculations (job of theorists), we require validated statistical models from the experiments!
- Expected backgrounds not enough: need full covariance information
- See e.g. CMS simplified likelihoods*:
    - CMS-NOTE-2017-001, *"Simplified likelihood for the re-interpretation of public CMS results"*



$$\mathcal{L}_S(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta}\right)$$

*Actually we need pdfs, not just likelihood function shapes, but in this case the pdf is clear.

If no covariance information published?

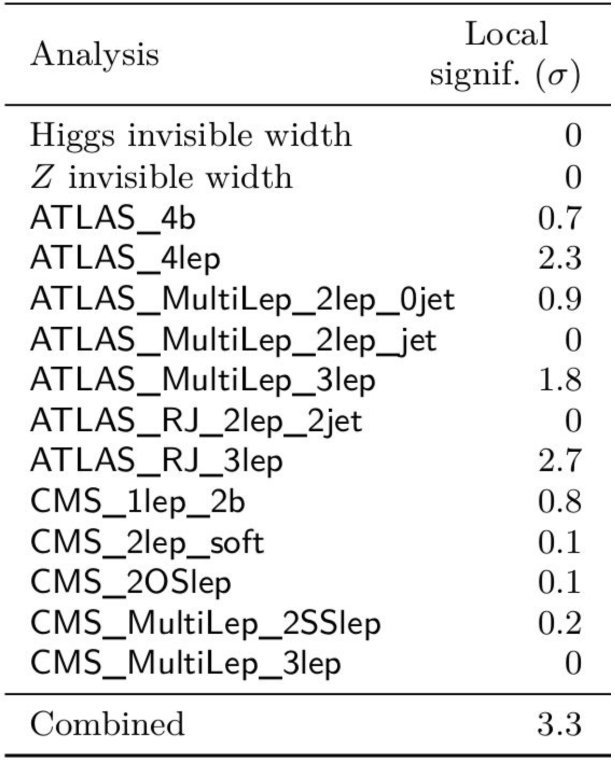
Two possibilities:

1. Ignore correlations (assume SRs are independent)
   a. In many cases may not be a terrible assumption. But no guarantees.
2. Pre-select “best expected” SR, and use only that one.
   a. Generally conservative (in the sense that will exclude fewer signal models, usually)
   b. Sensitive to choice of SR, throws out lots of information in other SRs!

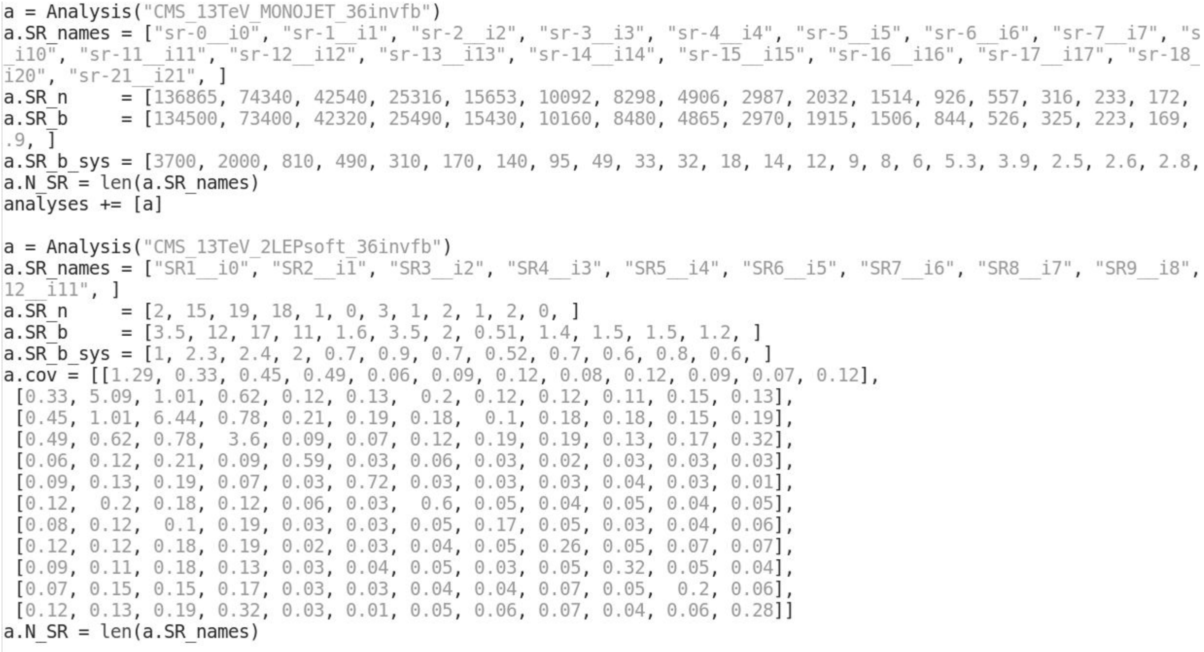
Much better to have the correlation information, even approximately!

(Regarding 2: this probably also messes up Wilks’ theorem; different parts of parameter space have different SRs pre-selected, so sudden jumps exist in likelihood surface -> not differentiable!)

Example:



Analysis	Local signif. (σ)
Higgs invisible width	0
Z invisible width	0
ATLAS_4b	0.7
ATLAS_4lep	2.3
ATLAS_MultiLep_2lep_0jet	0.9
ATLAS_MultiLep_2lep_jet	0
ATLAS_MultiLep_3lep	1.8
ATLAS_RJ_2lep_2jet	0
ATLAS_RJ_3lep	2.7
CMS_1lep_2b	0.8
CMS_2lep_soft	0.1
CMS_2OSlep	0.1
CMS_MultiLep_2SSlep	0.2
CMS_MultiLep_3lep	0
Combined	3.3

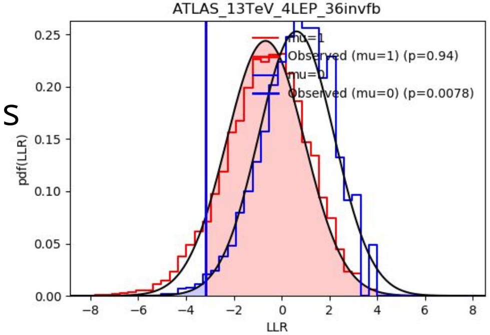


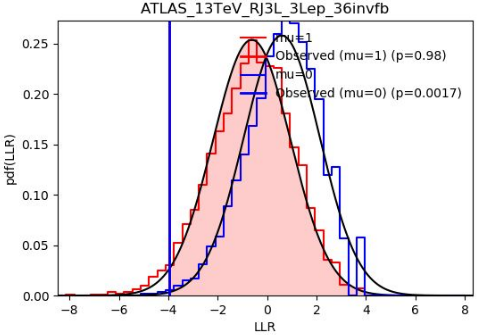
```
a = Analysis("CMS_13TeV_MONOJET_36invfb")
a.SR_names = ["sr-0__i0", "sr-1__i1", "sr-2__i2", "sr-3__i3", "sr-4__i4", "sr-5__i5", "sr-6__i6", "sr-7__i7", "s
_i10", "sr-11__i11", "sr-12__i12", "sr-13__i13", "sr-14__i14", "sr-15__i15", "sr-16__i16", "sr-17__i17", "sr-18_
i20", "sr-21__i21", ]
a.SR_n     = [136865, 74340, 42540, 25316, 15653, 10092, 8298, 4906, 2987, 2032, 1514, 926, 557, 316, 233, 172,
a.SR_b     = [134500, 73400, 42320, 25490, 15430, 10160, 8480, 4865, 2970, 1915, 1506, 844, 526, 325, 223, 169,
.9, ]
a.SR_b_sys = [3700, 2000, 810, 490, 310, 170, 140, 95, 49, 33, 32, 18, 14, 12, 9, 8, 6, 5.3, 3.9, 2.5, 2.6, 2.8,
a.N_SR = len(a.SR_names)
analyses += [a]

a = Analysis("CMS_13TeV_2LEPsoft_36invfb")
a.SR_names = ["SR1__i0", "SR2__i1", "SR3__i2", "SR4__i3", "SR5__i4", "SR6__i5", "SR7__i6", "SR8__i7", "SR9__i8",
12__i11", ]
a.SR_n     = [2, 15, 19, 18, 1, 0, 3, 1, 2, 1, 2, 0, ]
a.SR_b     = [3.5, 12, 17, 11, 1.6, 3.5, 2, 0.51, 1.4, 1.5, 1.5, 1.2, ]
a.SR_b_sys = [1, 2.3, 2.4, 2, 0.7, 0.9, 0.7, 0.52, 0.7, 0.6, 0.8, 0.6, ]
a.cov = [[1.29, 0.33, 0.45, 0.49, 0.06, 0.09, 0.12, 0.08, 0.12, 0.09, 0.07, 0.12],
 [0.33, 5.09, 1.01, 0.62, 0.12, 0.13,  0.2, 0.12, 0.12, 0.11, 0.15, 0.13],
 [0.45, 1.01, 6.44, 0.78, 0.21, 0.19, 0.18,  0.1, 0.18, 0.18, 0.15, 0.19],
 [0.49, 0.62, 0.78,  3.6, 0.09, 0.07, 0.12, 0.19, 0.19, 0.13, 0.17, 0.32],
 [0.06, 0.12, 0.21, 0.09, 0.59, 0.03, 0.06, 0.03, 0.02, 0.03, 0.03, 0.03],
 [0.09, 0.13, 0.19, 0.07, 0.03, 0.72, 0.03, 0.03, 0.03, 0.04, 0.03, 0.01],
 [0.12,  0.2, 0.18, 0.12, 0.06, 0.03,  0.6, 0.05, 0.04, 0.05, 0.04, 0.05],
 [0.08, 0.12,  0.1, 0.19, 0.03, 0.03, 0.05, 0.17, 0.05, 0.03, 0.04, 0.06],
 [0.12, 0.12, 0.18, 0.19, 0.02, 0.03, 0.04, 0.05, 0.26, 0.05, 0.07, 0.07],
 [0.09, 0.11, 0.18, 0.13, 0.03, 0.04, 0.05, 0.03, 0.05, 0.32, 0.05, 0.04],
 [0.07, 0.15, 0.15, 0.17, 0.03, 0.03, 0.04, 0.04, 0.07, 0.05,  0.2, 0.06],
 [0.12, 0.13, 0.19, 0.32, 0.03, 0.01, 0.05, 0.06, 0.07, 0.04, 0.06, 0.28]]
a.N_SR = len(a.SR_names)
```

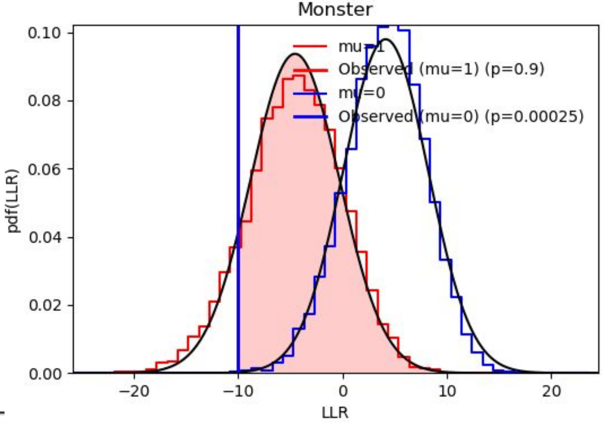
arXiv:1809.0209, *“Combined collider constraints on neutralinos and charginos“*

Example:

Analysis	BE SR Local signif. (σ)	No correlations Local signif. (σ)
Higgs invisible width	0	0
Z invisible width	0	0
ATLAS_4b	0.7	1.5
ATLAS_4lep	2.3	2.5
ATLAS_MultiLep_2lep_0jet	0.9	1.3
ATLAS_MultiLep_2lep_jet	0	0.8
ATLAS_MultiLep_3lep	1.8	1.2
ATLAS_RJ_2lep_2jet	0	1.5
ATLAS_RJ_3lep	2.7	3.4
CMS_1lep_2b	0.8	0
CMS_2lep_soft	0.1	0.1
CMS_2OSlep	0.1	0
CMS_MultiLep_2SSlep	0.2	0.2
CMS_MultiLep_3lep	0	0
Combined	3.3	4.1







$$q = -2\ln\frac{L_{s+b}}{L_b}\ ,$$

$$q = -2\ln\frac{L(\mu=1,\hat{\hat{\boldsymbol{\theta}}}(1))}{L(\mu=0,\hat{\hat{\boldsymbol{\theta}}}(0))}$$

arXiv:1809.0209, *“Combined collider constraints on neutralinos and charginos“*

Model independent approach?

Can we do this in a model-independent way?

Kind of: assign a free parameter to every signal region in every analysis!

$$\lambda_i = \mu s_i + b_i$$

Ok if not too many signal regions.

But power of test is quickly destroyed if there are many signal regions*

*Pretty much due to “look elsewhere effect” correction

Different test statistic required, since alternate hypothesis is not fixed

→ but this parameter space has good regularity properties, so maximum likelihood ratio test statistic is a good choice:

$$q_{\mathrm{GOF}} = -2\log\frac{\mathcal{L}_{\mathrm{joint}}(\mathbf{s}(\theta), \hat{\eta})}{\mathcal{L}_{\mathrm{joint}}(\hat{\hat{\mathbf{s}}}, \hat{\hat{\eta}})}$$

Asymptotic distribution is chi-squared with DOF=#SR

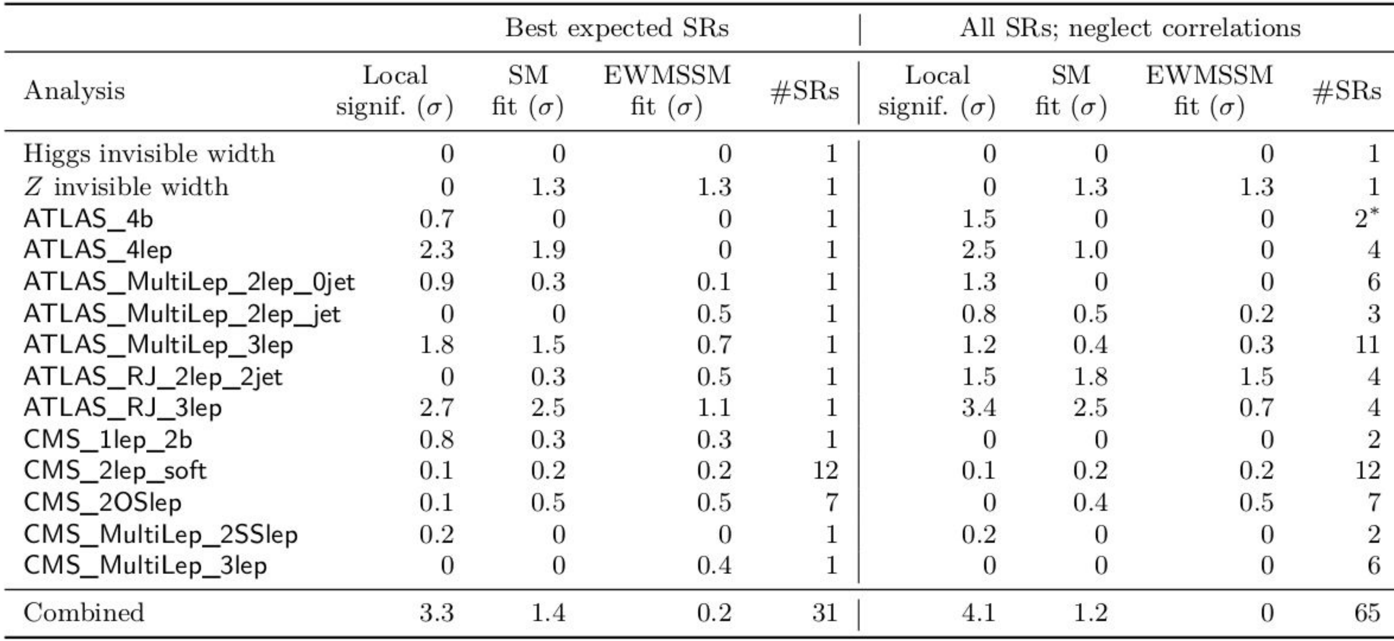
(to test background-only hypothesis, set s(θ)=0)

- No look-elsewhere problem here, have a super general signal model

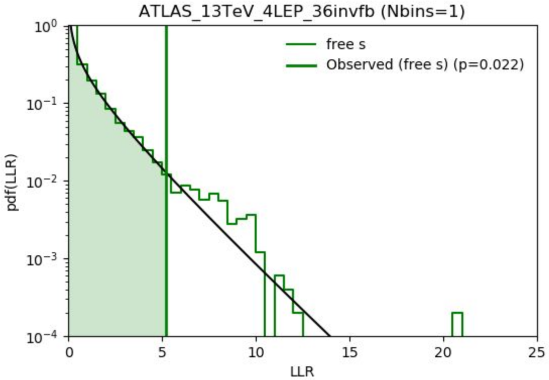
… Actually, too general! More freedom than in any real theory! This reduces the discover power more than is correct.

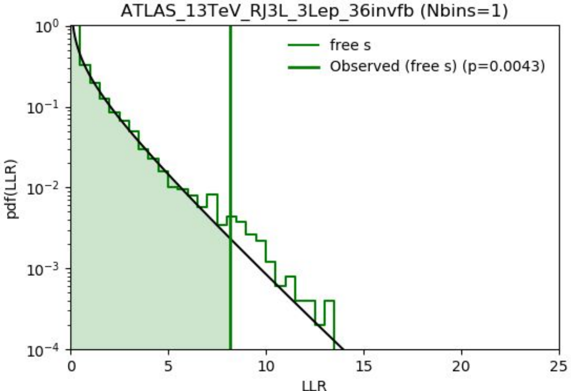
- Actually, look-elsewhere problem not totally gone if “best-expected signal region” method used; then we only have one parameter, but many possible signal regions that could be selected as “best” for different signal models.
    - (Another reason we need the correlation information!)

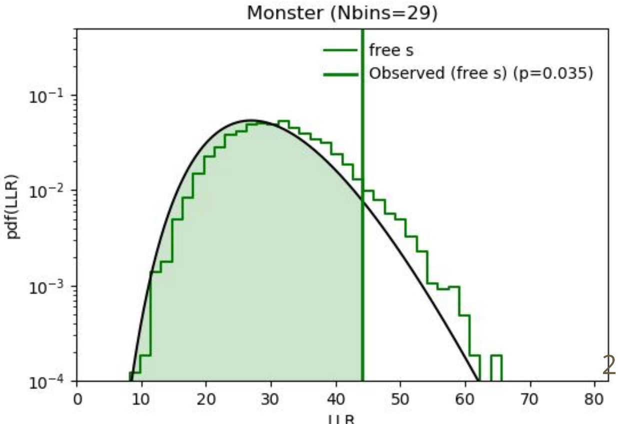
Example:



	Best expected SRs				All SRs; neglect correlations			
Analysis	Local signif. (σ)	SM fit (σ)	EWMSSM fit (σ)	#SRs	Local signif. (σ)	SM fit (σ)	EWMSSM fit (σ)	#SRs
Higgs invisible width	0	0	0	1	0	0	0	1
Z invisible width	0	1.3	1.3	1	0	1.3	1.3	1
ATLAS_4b	0.7	0	0	1	1.5	0	0	2^*
ATLAS_4lep	2.3	1.9	0	1	2.5	1.0	0	4
ATLAS_MultiLep_2lep_0jet	0.9	0.3	0.1	1	1.3	0	0	6
ATLAS_MultiLep_2lep_jet	0	0	0.5	1	0.8	0.5	0.2	3
ATLAS_MultiLep_3lep	1.8	1.5	0.7	1	1.2	0.4	0.3	11
ATLAS_RJ_2lep_2jet	0	0.3	0.5	1	1.5	1.8	1.5	4
ATLAS_RJ_3lep	2.7	2.5	1.1	1	3.4	2.5	0.7	4
CMS_1lep_2b	0.8	0.3	0.3	1	0	0	0	2
CMS_2lep_soft	0.1	0.2	0.2	12	0.1	0.2	0.2	12
CMS_2OSlep	0.1	0.5	0.5	7	0	0.4	0.5	7
CMS_MultiLep_2SSlep	0.2	0	0	1	0.2	0	0	2
CMS_MultiLep_3lep	0	0	0.4	1	0	0	0	6
Combined	3.3	1.4	0.2	31	4.1	1.2	0	65







Problem:

- The “local significance” test is prone to false positives (no look-elsewhere correction)
- The “model independent” test is prone to false negatives (too much look-elsewhere correction)

The “correct” look-elsewhere correction depends on the space of signals that are under consideration!

For simple cases (e.g. Higgs search, 1 free parameter) can do this correction (see e.g. arxiv:1005.1891, “*Trial factors for the look elsewhere effect in high energy physics*”)

In simple parameter spaces with good regularity properties (similar to “model independent” test) can get away with asymptotic theory.

But in general parameter spaces it must be done numerically, and requires enormous CPU resources (need to do many MC simulations at points all across the parameter space!).

(locally-good asymptotic behaviour doesn’t really help when taking min. over par. space.)

→ New methods would be extremely welcome here!

Summary

- Local p-value construction useful
  - + Best discovery power
  - + Good asymptotic behaviour
  - - Will overstate true significance of excesses (look elsewhere effect)

$$q = -2\ln \frac{L(\mu = 1, \hat{\hat{\boldsymbol{\theta}}}(1))}{L(\mu = 0, \hat{\hat{\boldsymbol{\theta}}}(0))}$$

- General signal parameterisation test semi-useful
  - + Look-elsewhere effect (mostly) removed
    - - ...but generally by too much! I.e. test often too weak, especially if lots of SRs
  - + Good asymptotic behaviour

$$q_{\mathrm{GOF}} = -2\log \frac{\mathcal{L}_{\mathrm{joint}}(\mathbf{s}(\theta), \hat{\eta})}{\mathcal{L}_{\mathrm{joint}}(\hat{\hat{\mathbf{s}}}, \hat{\hat{\eta}})}$$

- Super expensive to “brute force” the correct LEE correction! Is there a clever way to do it?
- Not discussed:
  - Correlations between analyses (i.e. common nuisance parameters)
  - Hopefully not a big effect (if orthogonal events selected), but currently no way to account for this (need full statistical model, i.e. likelihood, from experiments)
    - ...unless there is some clever “next-to-simplified” likelihood that can be published?
  - Various approximations required, so the significance of any interesting excesses will ultimately need to be confirmed by experiments.