HepData for BSM reinterpretation of LHC data

Andy Buckley, University of Glasgow

HepData Advisory Meeting, 24 Nov 2017

Intro

- **Quantity and quality of HepData content from LHC has been steadily improving – including ad hoc "auxiliary data"**
- BSM pheno community very happy about this – *vision of an automated limit re-setting toolchain with comprehensive data coverage*
- Include both "SM measurements" & dedicated BSM search data
- Primary data faithfully recorded, modulo format details. Issue is with *secondary data*:
  - (MC) background estimates
  - Correlation data
- **All data needs to "automatically" flow from experiments, through HepData, and into analysis tools**
  $\Rightarrow$ standardise formats and conventions for data & aux data

Intro

- **Quantity and quality of HepData content from LHC has been steadily improving – including ad hoc "auxiliary data"**
- BSM pheno community very happy about this – *vision of an automated limit re-setting toolchain with comprehensive data coverage*
- Include both "SM measurements" & dedicated BSM search data
- Primary data faithfully recorded, modulo format details. Issue is with *secondary data*:
  - (MC) background estimates
  - Correlation data
- **All data needs to "automatically" flow from experiments, through HepData, and into analysis tools**
  $\Rightarrow$ standardise formats and conventions for data & aux data
- **Frames a potential work-plan to include in funding application**

Correlations in fits/limit setting

Many types of correlation:

- **Between bins/SRs,** introduced by experimental/theory systematics
- **Between bins/analyses,** introduced by sharing events (or normalisation)
- **Between systematic (nuisance) params,** induced by profile fitting

Correlations in fits/limit setting

Many types of correlation:

- **Between bins/SRs,** introduced by experimental/theory systematics
- **Between bins/analyses,** introduced by sharing events (or normalisation)
- **Between systematic (nuisance) params,** induced by profile fitting

Possible approaches to providing this information:

- **full likelihood expression**, e.g. HistFactory demo
- approximate: **express as independent error sources**, correlated across bins — *extensible*
- approximate: **simplified likelihoods**: drop connection to error sources, bkg systs only, express as (symm) bin covariance
  Actively used by CMS: https://cds.cern.ch/record/2242860

Correlations in fits/limit setting

Many types of correlation:

- **Between bins/SRs,** introduced by experimental/theory systematics
- **Between bins/analyses,** introduced by sharing events (or normalisation)
- **Between systematic (nuisance) params,** induced by profile fitting

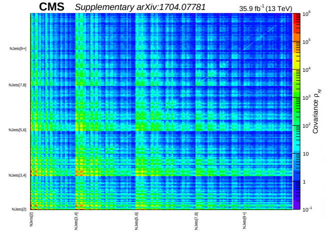
Possible approaches to providing this information:

- **full likelihood expression**, e.g. HistFactory demo
- approximate: **express as independent error sources**, correlated across bins — *extensible*
- approximate: **simplified likelihoods**: drop connection to error sources, bkg systs only, express as (symm) bin covariance
  Actively used by CMS: https://cds.cern.ch/record/2242860

Not 100% clear that correlations are necessary, but without them there will always be questions of whether an analysis was too optimistic or conservative

Correlation formats: error sources vs. bin covariance

CMS 0ℓ cov matrix – note log-scale!



Error breakdown in a HepData record

NB. normal in *Standard Model* analyses

RE	P P --> JETS
COS PHI	TEEC
-1 - -0.96	10.5165 ±0.00779481 stat +0.0117651 -0.0113337 sys,jesNp1 +0.0034308 -0.00335944 sys,jesNp2 + 71 more errors Show all
-0.96 - -0.92	0.716955 ±0.00468718 stat +0.00357606 -0.00426249 sys,jesNp1 +0.00165822 -0.00199598 sys,jesNp2 + 71 more errors Show all
-0.92 - -0.88	0.322052 ±0.00299636 stat +0.00184137 -0.00189796 sys,jesNp1 +0.000834961 -0.00104189 sys,jesNp2 + 71 more errors Show all

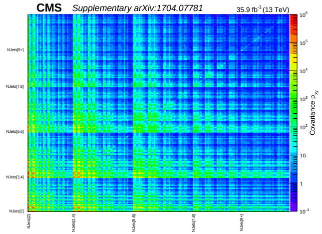
SL originally formalised as symm covariance

Simple to use: $L(\mu, \vec{\theta}) = \prod_i \mathrm{Pois}(n_i, \mu, \vec{\theta}) \cdot \mathrm{Gaus}(\vec{\theta}, \boldsymbol{C})$

Dimensionality of cov fixed: uniform approach, scales well. But limited to symmetric errs and no correlations between analyses.

Correlation formats: error sources vs. bin covariance

CMS 0ℓ cov matrix – note log-scale!



Error breakdown in a HepData record

NB. normal in *Standard Model* analyses

RE	P P --> JETS
COS PHI	TEEC
-1 - -0.96	10.5165 ±0.00779481 stat +0.0117651 -0.0113337 sys,jesNp1 +0.0034308 -0.00335944 sys,jesNp2 + 71 more errors Show all
-0.96 - -0.92	0.716955 ±0.00468718 stat +0.00357606 -0.00426249 sys,jesNp1 +0.00165822 -0.00199598 sys,jesNp2 + 71 more errors Show all
-0.92 - -0.88	0.322052 ±0.00299636 stat +0.00184137 -0.00189796 sys,jesNp1 +0.000834961 -0.00104189 sys,jesNp2 + 71 more errors Show all

SL originally formalised as symm covariance

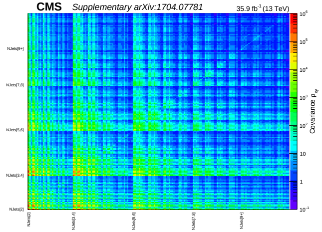
Simple to use: $L(\mu, \vec{\theta}) = \prod_i \mathrm{Pois}(n_i, \mu, \vec{\theta}) \cdot \mathrm{Gaus}(\vec{\theta}, \boldsymbol{C})$

Dimensionality of cov fixed: uniform approach, scales well. But limited to symmetric errs and no correlations between analyses.

HepData doesn't understand datasets semantics: would need "link" metadata to reliably connect correlation datasets to primary datasets

Correlation formats: error sources vs. bin covariance

CMS 0ℓ cov matrix – note log-scale!



Error breakdown in a HepData record
NB. normal in *Standard Model* analyses

RE	P P --> JETS
COS PHI	TEEC
-1 - -0.96	10.5165 ±0.00779481 stat; +0.0117651 -0.0113337 sys,jesNp1; +0.0034308 -0.00335944 sys,jesNp2; + 71 more errors Show all
-0.96 - -0.92	0.716955 ±0.00468718 stat; +0.00357606 -0.00426249 sys,jesNp1; +0.00165822 -0.00199598 sys,jesNp2; + 71 more errors Show all
-0.92 - -0.88	0.322052 ±0.00299636 stat; +0.00184137 -0.00189796 sys,jesNp1; +0.000834961 -0.00104189 sys,jesNp2; + 71 more errors Show all

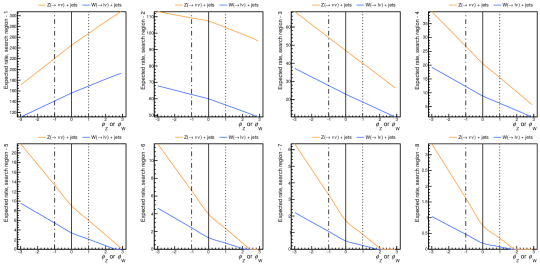
SL originally formalised as symm covariance
Simple to use: $L(\mu, \vec{\theta}) = \prod_i \mathrm{Pois}(n_i, \mu, \vec{\theta}) \cdot \mathrm{Gaus}(\vec{\theta}, \boldsymbol{C})$
Dimensionality of cov fixed: uniform approach, scales well. But limited to symmetric errs and no correlations between analyses.

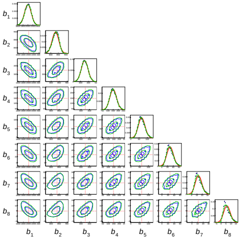
HepData doesn't understand datasets semantics: would need "link" metadata to reliably connect correlation datasets to primary datasets

Error-source representation more flexible: can construct cov matrix $C_{ij} = \sum_e \sigma_i \sigma_j$, or asymm by toy-sampling
Extensible! Supported already. HD preference. But…

Logistical issues & extensions

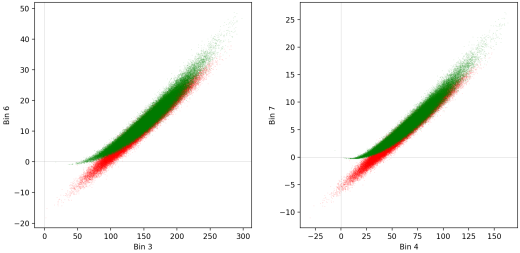
- Need standard names, esp. to distinguish uncorr stat errors
- Also need groupings, e.g. to separate theory/MC errors from experimental/detector resolutions ⇒ future reinterpretations with theory improvements. Easier with explicit cov matrices?
- Error-sources are naturally usable in an asymmetric way. But current activity on use of skew moments to implement asymm parametrisation: **how to store this in HD?!**





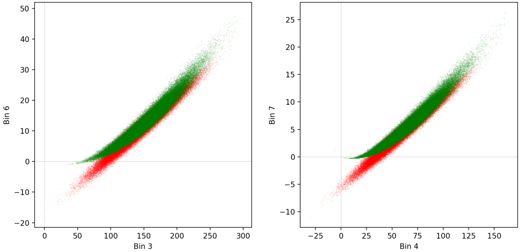
Logistical issues & extensions

- Need standard names, esp. to distinguish uncorr stat errors
- Also need groupings, e.g. to separate theory/MC errors from experimental/detector resolutions ⇒ future reinterpretations with theory improvements. Easier with explicit cov matrices?
- Error-sources are naturally usable in an asymmetric way. But current activity on use of skew moments to implement asymm parametrisation: **how to store this in HD?!**



Logistical issues & extensions

- Need standard names, esp. to distinguish uncorr stat errors
- Also need groupings, e.g. to separate theory/MC errors from experimental/detector resolutions ⇒ future reinterpretations with theory improvements. Easier with explicit cov matrices?
- Error-sources are naturally usable in an asymmetric way. But current activity on use of skew moments to implement asymm parametrisation: **how to store this in HD?!**



Possibility for HD to have *semantic* understanding of correlations?

MC and background data

- Correlations are the most technically complex demand, since the data objects are semantically different from “normal” datasets

MC and background data

- Correlations are the most technically complex demand, since the data objects are semantically different from “normal” datasets
- Not the only requirement for scalable recasting, though: background estimates are also crucial

MC and background data

- Correlations are the most technically complex demand, since the data objects are semantically different from "normal" datasets
- Not the only requirement for scalable recasting, though: background estimates are also crucial
- Typical BSM reinterpretations only have the capacity to generate (maybe LO) signal events

MC and background data

- Correlations are the most technically complex demand, since the data objects are semantically different from “normal” datasets
- Not the only requirement for scalable recasting, though: background estimates are also crucial
- Typical BSM reinterpretations only have the capacity to generate (maybe LO) signal events
- Backgrounds computed by experiments using vast MC datasets with very complex and CPU-intensive high-sophistication modelling: not reproducible, so needs to be published

MC and background data

- Correlations are the most technically complex demand, since the data objects are semantically different from "normal" datasets
- Not the only requirement for scalable recasting, though: background estimates are also crucial
- Typical BSM reinterpretations only have the capacity to generate (maybe LO) signal events
- Backgrounds computed by experiments using vast MC datasets with very complex and CPU-intensive high-sophistication modelling: not reproducible, so needs to be published
- This has started, but – again – **how to make HD (and its API) *semantically* aware of what is data and what's the corresponding MC?**

  And background process breakdown? And pre-/post-fit? …