Standard communication of analysis data correlations

Andy Buckley, University of Glasgow

LHC EW WG meeting, 22 June 2018

Intro

- **Rivet and similar preservation, tuning, recasting and other reinterpretation tools rely on HepData**

Intro

- **Rivet and similar preservation, tuning, recasting and other reinterpretation tools rely on HepData**

- Quantity and quality of HepData content has steadily improved through LHC – including ad hoc "auxiliary" data
  $\Rightarrow$ *coherent reinterpretation toolchains with full LHC data coverage*

Intro

- **Rivet and similar preservation, tuning, recasting and other reinterpretation tools rely on HepData**
- Quantity and quality of HepData content has steadily improved through LHC – including ad hoc "auxiliary" data
  $\Rightarrow$ *coherent reinterpretation toolchains with full LHC data coverage*
- Primary data faithfully recorded, modulo format details. Issue is with *secondary data*
  - Correlation data
  - For searches: (MC) background estimates

  LHC maturity and large integrated lumi $\Rightarrow$ a major issue for use of both measurement and search analysis data

Intro

- **Rivet and similar preservation, tuning, recasting and other reinterpretation tools rely on HepData**
- Quantity and quality of HepData content has steadily improved through LHC – including ad hoc "auxiliary" data
  $\Rightarrow$ *coherent reinterpretation toolchains with full LHC data coverage*
- Primary data faithfully recorded, modulo format details. Issue is with *secondary data*
  - Correlation data
  - For searches: (MC) background estimates

  LHC maturity and large integrated lumi $\Rightarrow$ a major issue for use of both measurement and search analysis data
- **All data needs to "automatically" flow from experiments $\rightarrow$ HepData $\rightarrow$ analysis tools**
  $\Rightarrow$ standardise formats and conventions for data & aux data
  $\Rightarrow$ *using* correlations in MC tuning, EFT fits, BSM scans = "easy"
  (formats like Rivet's YODA being extended to handle this: correlations via metadata & multiple named error bars)

Correlations in fits/limit setting

Many types of correlation:

- **SYST: between bins/SRs,** from experiment/theory systematics
- **STAT: between bins/analyses**, from event-sharing/normalisation
- **FIT: between systematic (nuisance) params**, via profile fitting

Correlations in fits/limit setting

Many types of correlation:

- **SYST: between bins/SRs,** from experiment/theory systematics
- **STAT: between bins**/analyses, from event-sharing/normalisation
- **FIT: between systematic (nuisance) params**, via profile fitting

Possible approaches to providing this information:

- **full likelihood expression**, e.g. HistFactory demo
  recent development of `pyhf` JSON format interesting…also usable by other tools? Likelihood generalisation required

Correlations in fits/limit setting

Many types of correlation:

- **SYST: between bins/SRs,** from experiment/theory systematics
- **STAT: between bins/analyses**, from event-sharing/normalisation
- **FIT: between systematic (nuisance) params**, via profile fitting

Possible approaches to providing this information:

- **full likelihood expression**, e.g. HistFactory demo
  recent development of `pyhf` JSON format interesting…also usable by other tools? Likelihood generalisation required
- **approx: leading correlation moments**, e.g. covariance matrix
  - correlated across bins, including between distributions
  - simplified likelihoods : drop connection to elementary error sources (opacity = useful?!) $\Rightarrow$ covariance sufficient?
  - extension beyond covariance possible but awkward

Correlations in fits/limit setting

Many types of correlation:

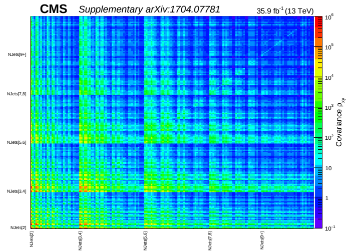
- **SYST: between bins/SRs,** from experiment/theory systematics
- **STAT: between bins**/analyses, from event-sharing/normalisation
- **FIT: between systematic (nuisance) params**, via profile fitting

Possible approaches to providing this information:

- **full likelihood expression**, e.g. HistFactory demo
  recent development of `pyhf` JSON format interesting…also usable by other tools? Likelihood generalisation required
- **approx: leading correlation moments**, e.g. covariance matrix
  - correlated across bins, including between distributions
  - simplified likelihoods : drop connection to elementary error sources (opacity = useful?!) $\Rightarrow$ covariance sufficient?
  - extension beyond covariance possible but awkward
- **representation options:** independent error sources on data points; linked primary/secondary datasets; `pyhf`

Correlation formats: error sources vs. bin covariance

CMS 0ℓ cov matrix (log-scale!)



Error breakdown in a HepData record

NB. normal in *Standard Model* analyses

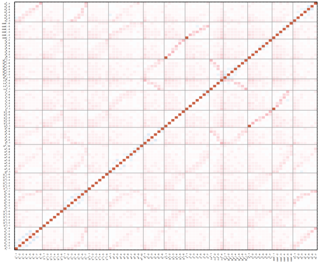
RE	P P --> JETS
COS PHI	**TEEC**
-1 - -0.96	10.5165 ±0.00779481 stat +0.0117651 -0.0113337 sys,jesNp1 +0.0034308 -0.00335944 sys,jesNp2 + 71 more errors Show all
-0.96 - -0.92	0.716955 ±0.00468718 stat +0.00357606 -0.00426249 sys,jesNp1 +0.00165822 -0.00199598 sys,jesNp2 + 71 more errors Show all
-0.92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

Covariance/correlation matrix as a separate dataset/table is limited

- symmetric errs only
- awkward to map to primary distribution(s) bins

Correlation formats: error sources vs. bin covariance

ATLAS $t\bar{t}$ hadronic cov matrix



Error breakdown in a HepData record

NB. normal in *Standard Model* analyses

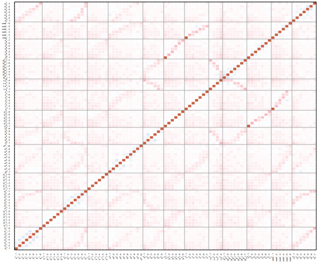
RE	P P --> JETS
COS PHI	**TEEC**
-1 - -0.96	10.5165 ±0.00779481 stat $^{+0.0117651}_{-0.0113337}$ sys,jesNp1 $^{+0.0034308}_{-0.00335944}$ sys,jesNp2 + 71 more errors Show all
-0.96 - -0.92	0.716955 ±0.00468718 stat $^{+0.00357606}_{-0.00426249}$ sys,jesNp1 $^{+0.00165822}_{-0.00199598}$ sys,jesNp2 + 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

Covariance/correlation matrix as a separate dataset/table is limited

- symmetric errs only
- awkward to map to primary distribution(s) bins

Correlation formats: error sources vs. bin covariance

ATLAS $t\bar{t}$ hadronic cov matrix



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

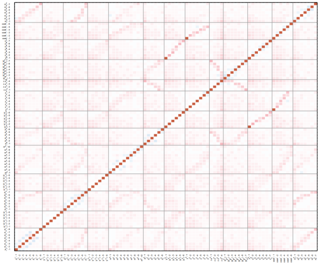
Covariance/correlation matrix as a separate dataset/table is limited

- symmetric errs only
- awkward to map to primary distribution(s) bins

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

Correlation formats: error sources vs. bin covariance

ATLAS $t\bar{t}$ hadronic cov matrix



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

Covariance/correlation matrix as a separate dataset/table is limited

- symmetric errs only
- awkward to map to primary distribution(s) bins

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

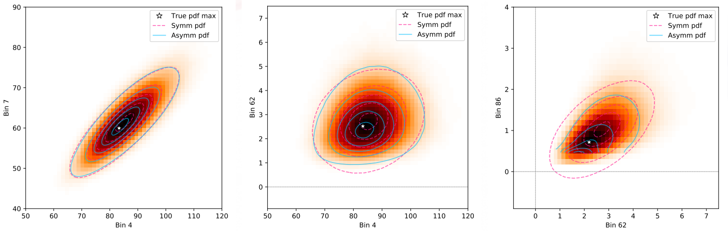
Error-source representation more flexible: can construct cov matrix $C_{ij} = \sum_e \sigma_i \sigma_j$, or asymm by toy-sampling

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
  $\Rightarrow$ future reinterpretations with theory improvements

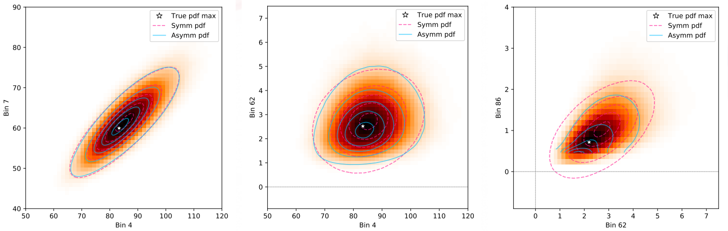
Logistical issues & extensions

- Need standard names, esp. to distinguish uncorr stat errors
- Also need groupings, e.g. to separate theory/MC errors from experimental/detector resolutions
  $\Rightarrow$ future reinterpretations with theory improvements
- 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
  $\Rightarrow$ future reinterpretations with theory improvements
- 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?!**



Extend HD for *semantic* correlation awareness? `pyhf` again...

Correlations as YODA metadata

YODA data format used by Rivet. Gradually extending data types to better match SM & BSM requirements. Input welcome…

Correlations as YODA metadata

YODA data format used by Rivet. Gradually extending data types to better match SM & BSM requirements. Input welcome. . .

Work by Louie Corpe & AB: auto-encode error source params from HepData as YODA histogram metadata

```
BEGIN YODA_SCATTER2D /ATLAS_2017_I1514251/d01-x06-y01
Corr: {0: {alphas: {dn: -0.02646259, up: 0.0003289776},
           norm: {dn: -0.1191564, up: 0.1191564},
           pdf: {dn: -0.02138033, up: 0.02138033},
           scale: {dn: -0.08166401, up: 0.04873643},
           stat: {dn: -0.01772649, up: 0.01772649}},
       1: { ...
```

Correlations as YODA metadata

YODA data format used by Rivet. Gradually extending data types to better match SM & BSM requirements. Input welcome…

Work by Louie Corpe & AB: auto-encode error source params from HepData as YODA histogram metadata

```
BEGIN YODA_SCATTER2D /ATLAS_2017_I1514251/d01-x06-y01
Corr: {0: {alphas: {dn: -0.02646259, up: 0.0003289776},
           norm: {dn: -0.1191564, up: 0.1191564},
           pdf: {dn: -0.02138033, up: 0.02138033},
           scale: {dn: -0.08166401, up: 0.04873643},
           stat: {dn: -0.01772649, up: 0.01772649}},
       1: { ...
```

Requires YAML-format headers in YODA: done in current release, modification to HepData export needed

Further work to support multiple errors on bins / data-points approaching release

Correlations as YODA metadata

YODA data format used by Rivet. Gradually extending data types to better match SM & BSM requirements. Input welcome…

Work by Louie Corpe & AB: auto-encode error source params from HepData as YODA histogram metadata

```
BEGIN YODA_SCATTER2D /ATLAS_2017_I1514251/d01-x06-y01
Corr: {0: {alphas: {dn: -0.02646259, up: 0.0003289776},
           norm: {dn: -0.1191564, up: 0.1191564},
           pdf: {dn: -0.02138033, up: 0.02138033},
           scale: {dn: -0.08166401, up: 0.04873643},
           stat: {dn: -0.01772649, up: 0.01772649}},
       1: { ...
```

Requires YAML-format headers in YODA: done in current release, modification to HepData export needed

Further work to support multiple errors on bins / data-points approaching release

What's the best way to propagate this info in a ROOT workflow?

Summary

- **Well advanced in many ways: correlations frequently assessed from several sources from systematics to fitting**

Summary

- **Well advanced in many ways: correlations frequently assessed from several sources from systematics to fitting**
- **Standardised reporting via HepData is key. Not all representations are equally good.**
  - "Matrix datasets" are least flexible
  - Error-source (+ skew?) or full likelihood are best. Some standard naming required
  - Integration of `pyhf` into HepData would allow for full semantic awareness of links between an analysis' primary & secondary datasets

Summary

- **Well advanced in many ways: correlations frequently assessed from several sources from systematics to fitting**
- **Standardised reporting via HepData is key. Not all representations are equally good.**
  - "Matrix datasets" are least flexible
  - Error-source (+ skew?) or full likelihood are best. Some standard naming required
  - Integration of `pyhf` into HepData would allow for full semantic awareness of links between an analysis' primary & secondary datasets
- **If we can access correlation data in a standard form, downstream tools will definitely use it**

Backup

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? ...