Machinery for providing and using correlation info

Andy Buckley, University of Glasgow

FNAL Reinterpretation Workshop, 16 Oct 2017

Intro

- Talk billed as "news from HepData" – but I can't speak for HD!
- Actually, biggest HD development has been in depth and quality of info provided there by expts!
- Instead, a short status update on the toolchain being assembled for correlation data propagation
- Corr data needs to "automatically" flow from experiments, through HepData, and into analysis tools
- $\Rightarrow$ Boring stuff like formats and conventions need to be $\sim$standardised!

Correlations in fits/limit setting

Soooo many types of correlation:

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

Correlations in fits/limit setting

Soooo many types of correlation:

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

In general, have to deal with a correlated joint pdf on nuisances, affecting bins/SRs in a correlated way

Possible approaches to providing this information:

- full likelihood expression → Lukas Heinrich HistFactory demo
- approximate: express as independent error sources, correlated across bins — *extensible*
- approximate: drop connection to error sources, bkg systs only, express as (symm) bin covariance ("simplified likelihoods ")

Correlations in fits/limit setting

Soooo many types of correlation:

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

In general, have to deal with a correlated joint pdf on nuisances, affecting bins/SRs in a correlated way

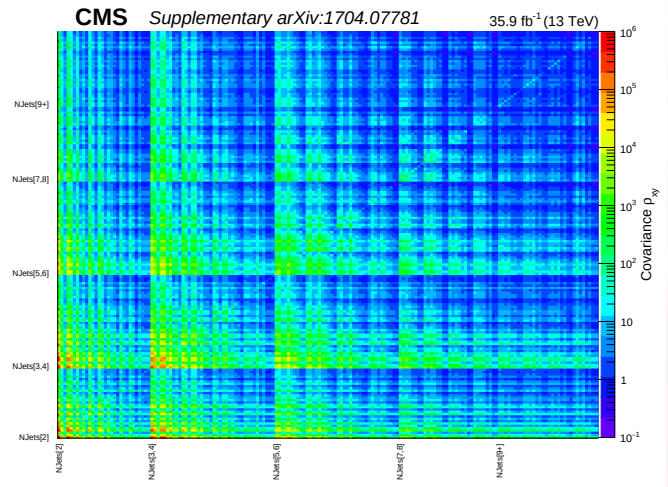
Possible approaches to providing this information:

- full likelihood expression → Lukas Heinrich HistFactory demo
- approximate: express as independent error sources, correlated across bins — *extensible*
- approximate: drop connection to error sources, bkg systs only, express as (symm) bin covariance ("simplified likelihoods ")

How ready are we?

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

Covariance simple to use: $L(\mu, \vec{\theta}) = \prod_i \text{Pois}(n_i, \mu, \vec{\theta}) \cdot \text{Gaus}(\vec{\theta}, \boldsymbol{C})$
Dimensionality of cov fixed: uniform approach, scales well. But ∼limited to symmetric errs and no correlations between analyses.

Error-source representation more flexible: can construct cov matrix $C_{ij} = \sum_e \sigma_i \sigma_j$, or asymm by toy-sampling. **Extensible!** But need standard names, esp. to distinguish diagonal stat errors

HepData doesn't understand datasets semantics: Add "link" metadata for covs? Cov matrix decomposition to error sources?

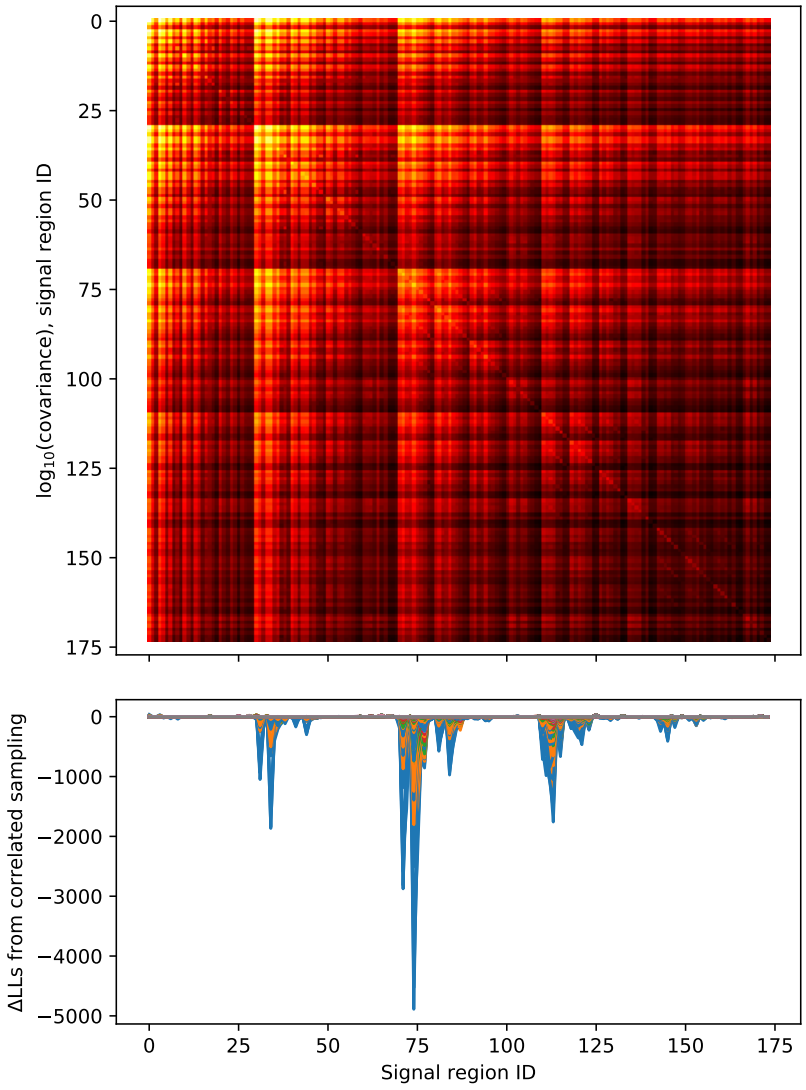
Example with SL bin covariance

Example of correlation (un)importance: CMS-SUS-16-033 0-lepton SUSY paper ↗ with 174 SRs & SL cov matrix

Corr data extraction quite manual.

Marginalise via multivariate normal sampling or n_{bin} 1D Gaussian integrations (cf. Gambit)

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



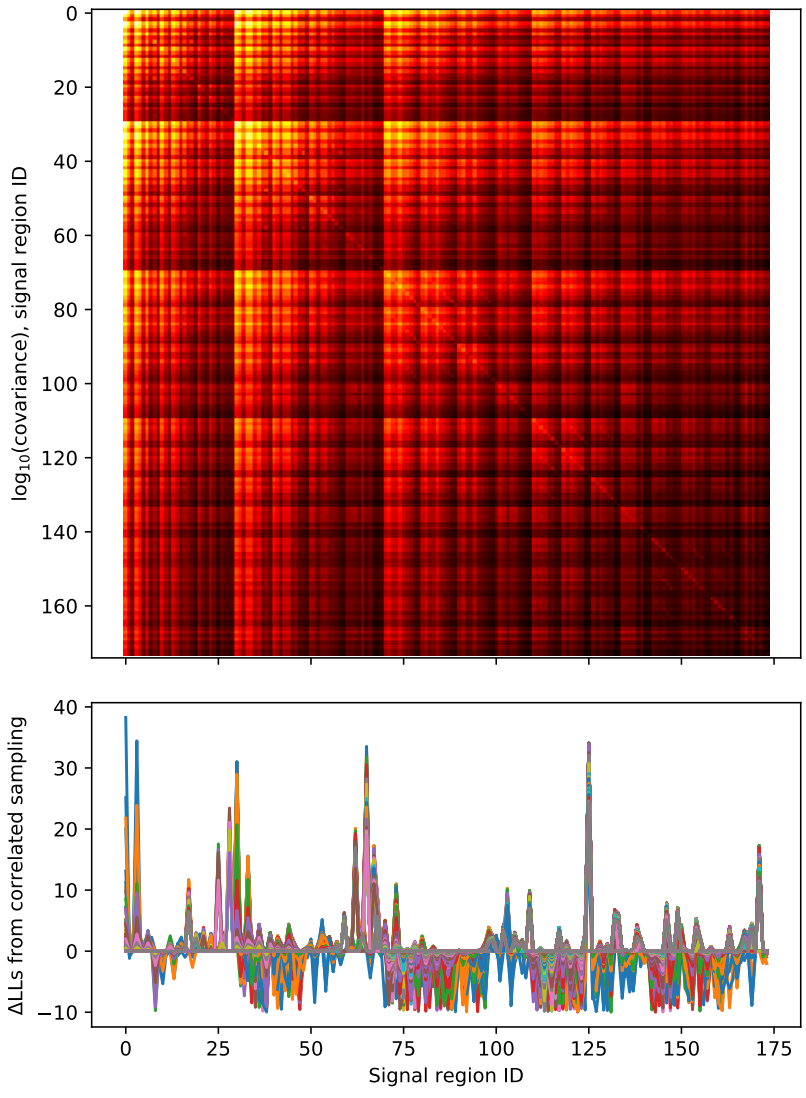
Example with SL bin covariance

Example of correlation (un)importance: CMS-SUS-16-033 0-lepton SUSY paper ↗ with 174 SRs & SL cov matrix

Corr data extraction quite manual.

Marginalise via multivariate normal sampling or n_{bin} 1D Gaussian integrations (cf. Gambit)

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



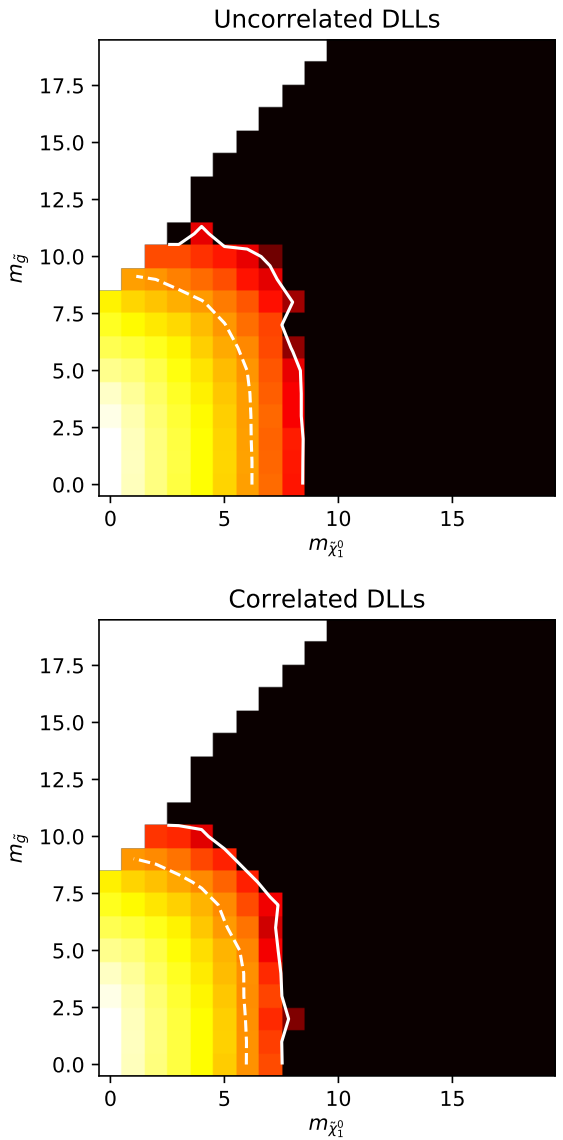
Example with SL bin covariance

Example of correlation (un)importance: CMS-SUS-16-033 0-lepton SUSY paper ↗ with 174 SRs & SL cov matrix

Corr data extraction quite manual.

Marginalise via multivariate normal sampling or n_{bin} 1D Gaussian integrations (cf. Gambit)

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



Correlations as YODA metadata

YODA data format used by Rivet. Gradually extending to a (minimal) set of data types sufficient for SM & BSM requirements. (Ideas and contributions welcome. . .)

Work by Louie Corpe: auto-encode error source params from HepData as YODA histogram metadata — then propagate by sampling (can be asymmetric) or by constructing covariance

```
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: work done, release imminent, modification to HepData export to follow shortly

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

Summary

- Several formalisms for correlation reporting & propagation, at various levels of sophistication
- General agreement that HepData should be the route for publishing this information. Needs support
- Small HD extensions for Rivet/YODA/Contur workflow coming — bigger action needed for metadata to express semantics, handle general likelihoods?
- No need for one formalism to "win", although some are more extensible than others. Further extensibility to allow incorporating improved theory errors as time passes
- Standards important $\Rightarrow$ common stats components?