Forward Folding

What? Why? How?

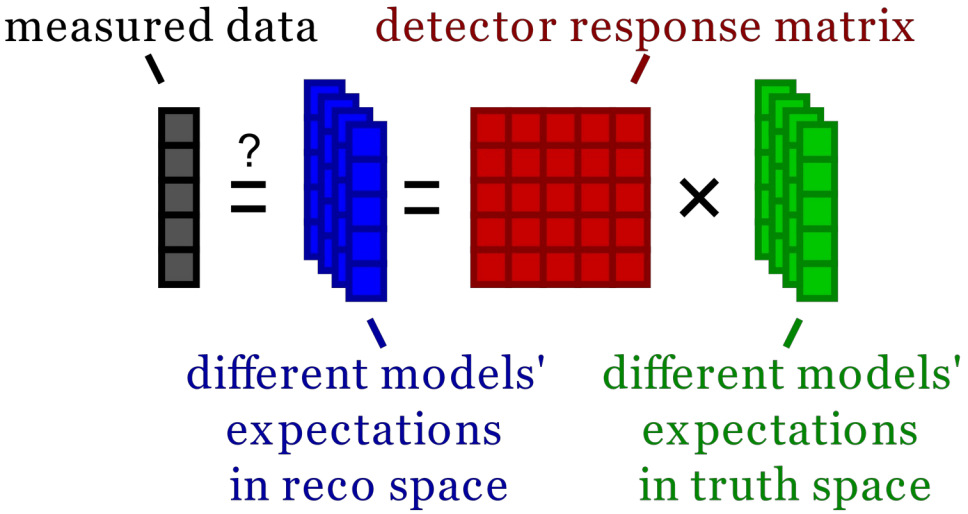
The cross-section cave

- We do not record the events as they happen
  - We record only what we are able to see
- What we see is not exactly what we are interested in
  - Lost events due to (in)efficiency
  - Added events due to background
  - Different event properties due to smearing

The cross-section cave

- The canonical way: Unfolding
  - “Undo” the detector and selection effects
  - Challenging to do right without introducing bias
  - Can be wrong in very subtle ways
- Another way: Forward-folding
  - Apply detector effects to model predictions
    - Response matrix
  - Brings its own set of challenges

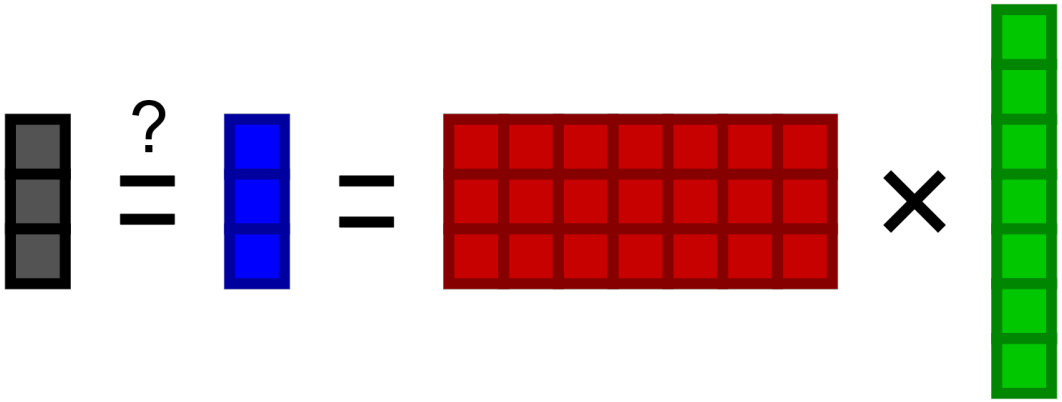
How it works



- Every event belongs in exactly one truth bin and up to one reconstructed bin (if it gets reconstructed)
- P(reco bin = i | truth bin = j) = R_{ij} = efficiency × smearing
  - Response matrix describes average detector response to true events
- reco expectation = response matrix × truth expectation
  - Can (and truth usually must) be binned in multiple variables

- The data is the data is the data
  - No uncertainty on the data points, 4 is exactly 4!
  - All systematics in response matrix or physics model
- All comparisons between data and theory (likelihoods, chi-squares, chi-by-eye) are done in reco space.
- Fast evaluations of many models possible

Flexible binning



- Flexible number of bins

  #reco bins != #truth bins

- Combine coarse reco binning with fine truth binning
- Good for
  - low statistics (need large signal MC sample though)
  - difficult to constrain efficiency variations
- Admit we are not able to constrain truth completely

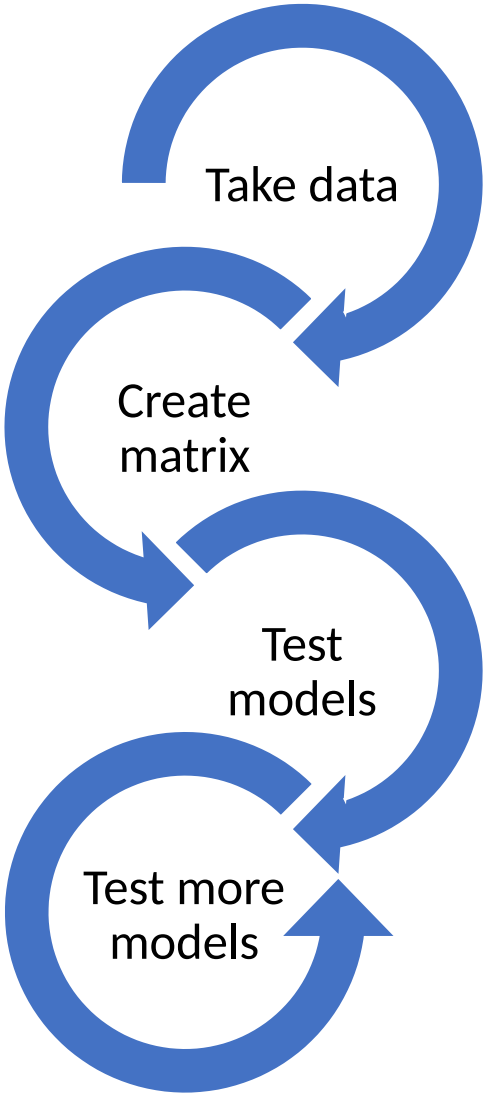
Reco level data

- No data point correlation
  - Theory predictions will be correlated, but probably much less than what unregularised unfolding might do
  - Chi-by-eye
- Robert D. Cousins, Samuel J. May, Yipeng Sun, "*Should unfolded histograms be used to test hypotheses?*": [arXiv:1607.07038]

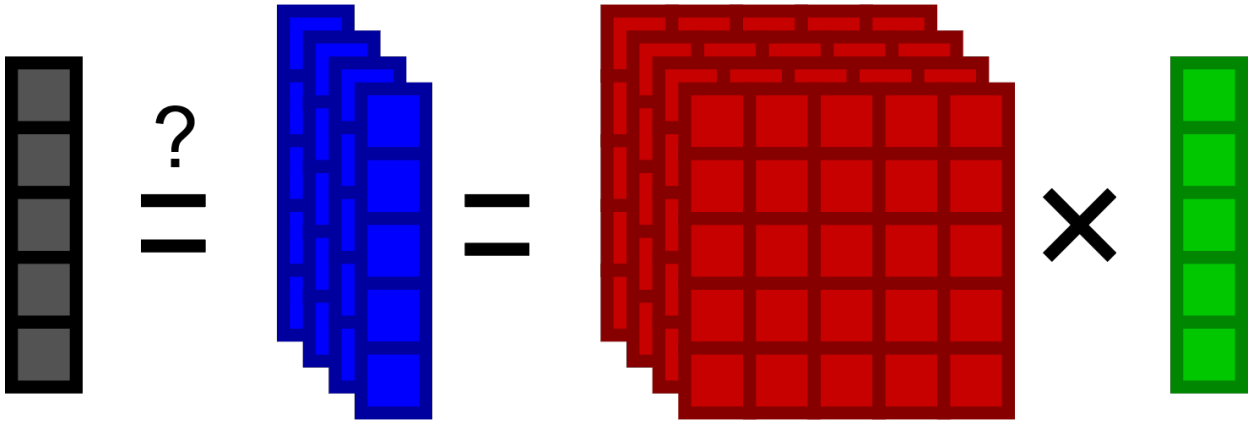
  "It seems remarkable that, even though unfolding by matrix inversion would appear not to lose information, in practice the way the information is used (linearizing the problem via expressing the result via a covariance matrix) already results in some failures of the bottom-line test of GOF. This is without any regularization or approximate EM inversion."

Data life cycle

- Publication of only data and matrix probably not feasible (yet)
- Can do model comparisons or parameter fit as part of result
- Raw data & model independent response matrix ensure maximum usefulness of data in the future
  - New models can be easily compared against old data
  - Including all detector systematics

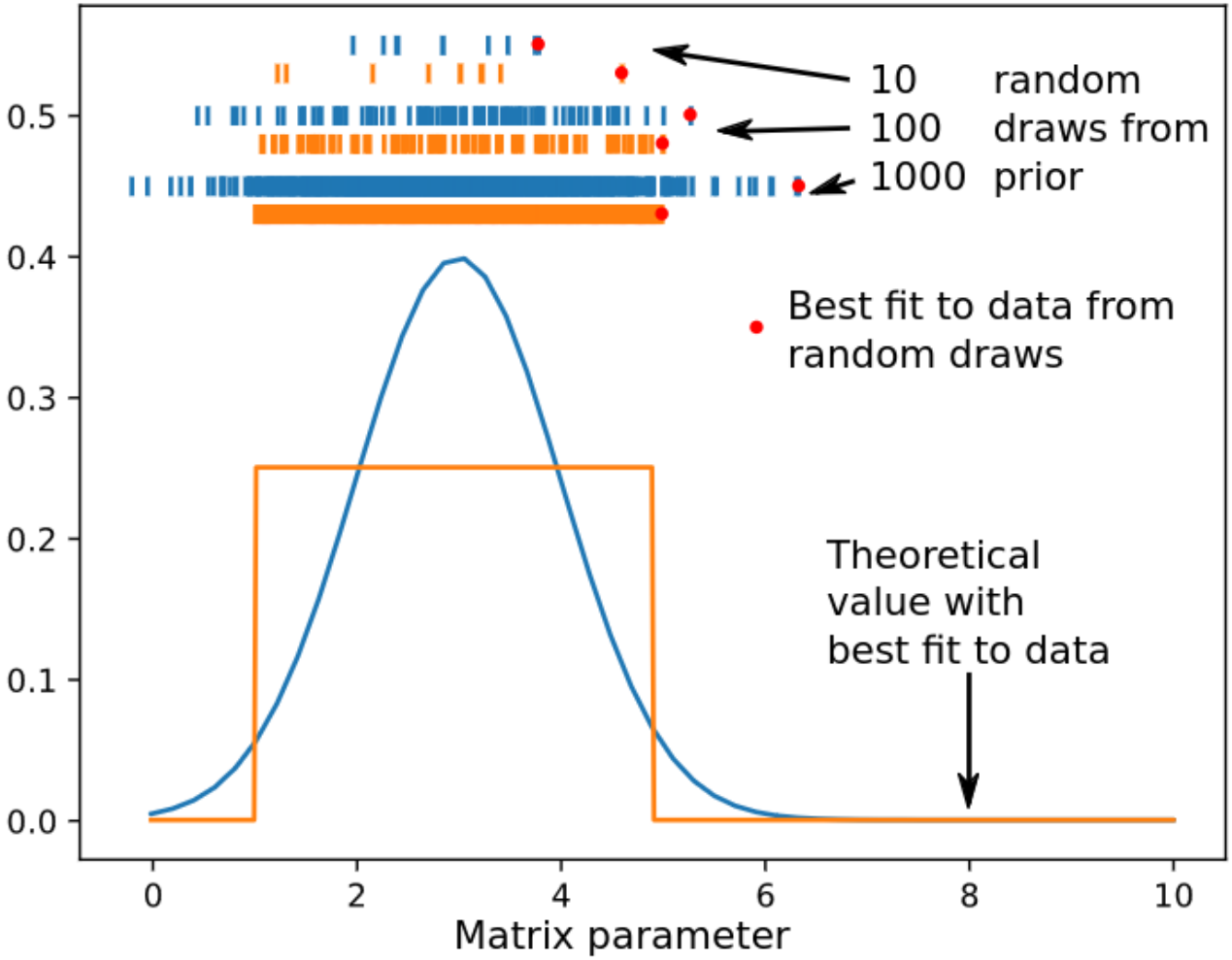


Detector uncertainties



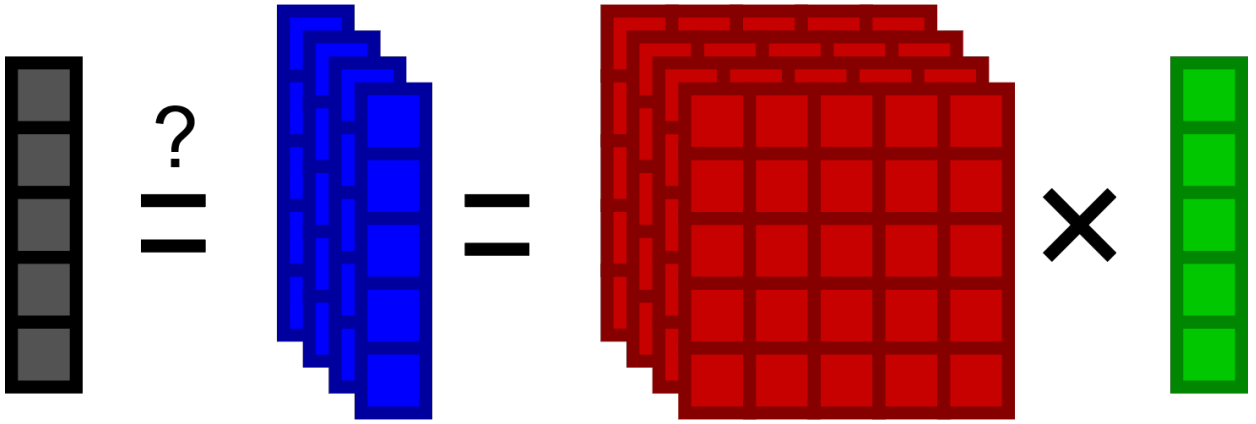
- One matrix only describes single possible detector
  - True detector probably behaves slightly differently
- Cover detector uncertainties with “toy simulations”
  - Variations and weights of same events
- Each toy yields own response matrix
- Each response matrix yields own reco prediction
- Compare to data w/ marginal, i.e. average, likelihood
  - No fitting of 1000 detector parameters!

Detector uncertainties



- One ma
  - True c
- Cover de
  - Variat
- Each toy
- Each res
- Compare to data w/ marginal, i.e. average, likelihood
  - No fitting of 1000 detector parameters!

Monte Carlo statistics

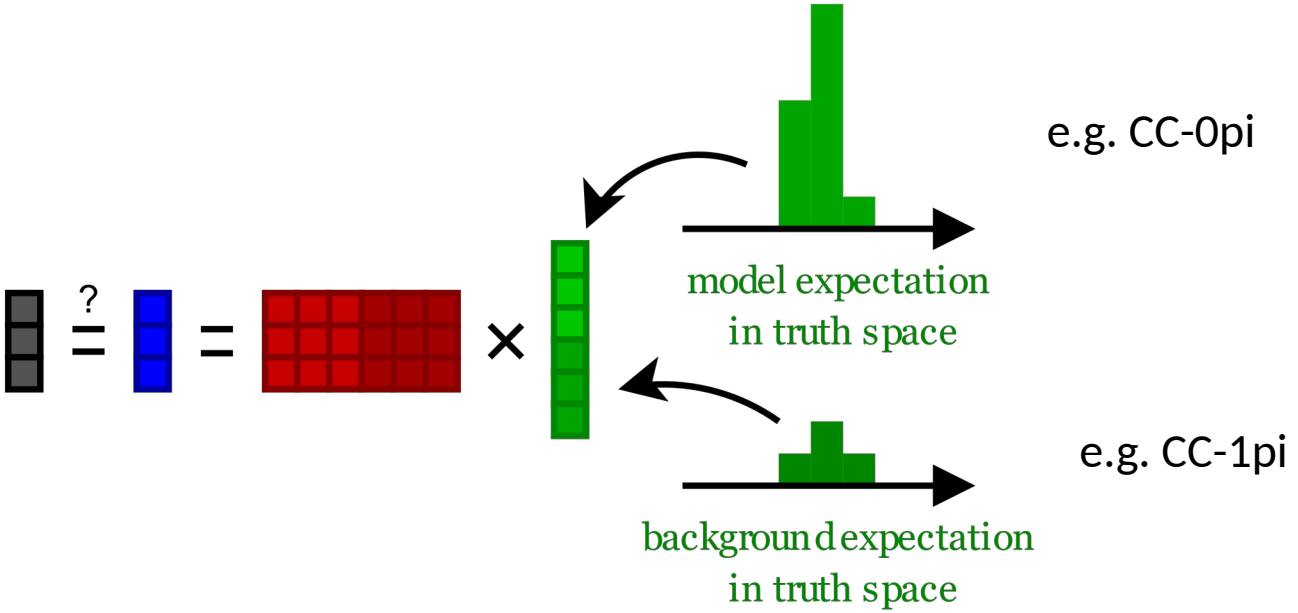


- Response matrix generated from MC event
- Limited statistics mean uncertainties on matrix elements
- Each column describes multinomial probabilites (including the “waste bin”)
  - R_{ij} = P(reco bin = i | truth bin = j)
- Can treat P as unknown Bayesian parameter with a probability denisty
- Described by Dirichlet distribution with hyperparmeters α_i from MC
- Sample from distirbution to get matrices randomly varied within uncertainties
- Treat like detector systematics

Background events

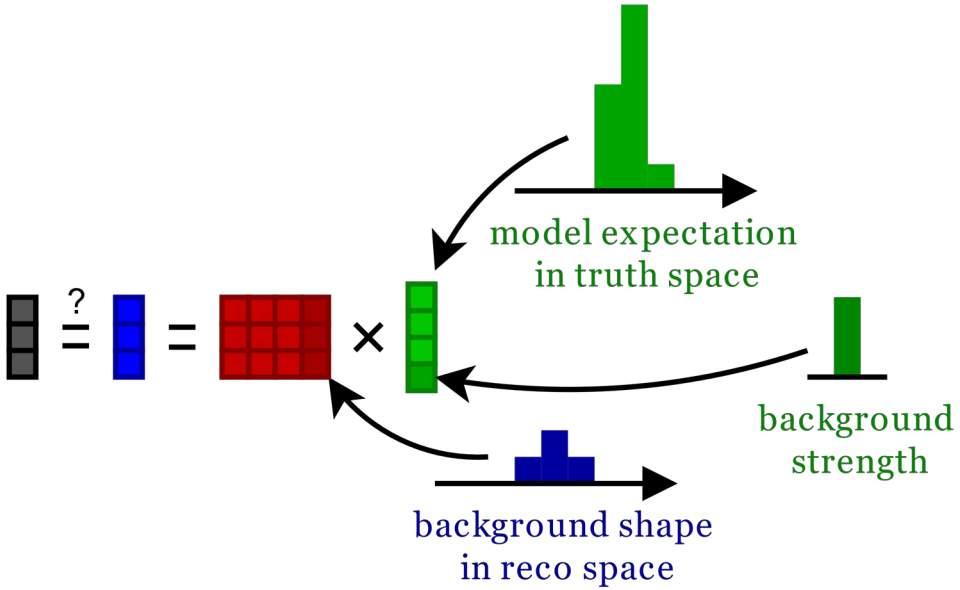
- Three kinds of backgrounds to distinguish:
- Irreducible background (same truth bins)
  - Events that are indistinguishable from signal on truth level
  - E.g. CC-RES with pi lost in FSI in CC-QE selection (Don’t do this)
  - Must be added to the respective truth bins
  - Can be constrained with control samples (model dependent!)
- Physics-like background (other truth bins)
  - Events with their own defined efficiency and smearing
  - E.g. CC-1pi events in CC-0pi selection
- Detector background (“no” truth bin)
  - Events where it is difficult to define true events
  - E.g. certain kinds of OOFV

Physics-like background



- Background where one can define a reasonable efficiency and smearing will be treated just like signal
- More columns in the response matrix
  - Equivalent to getting their own response matrix
- Users of the data might not care about those events
  - Provide templates or other simple models to go with the data

Detector background

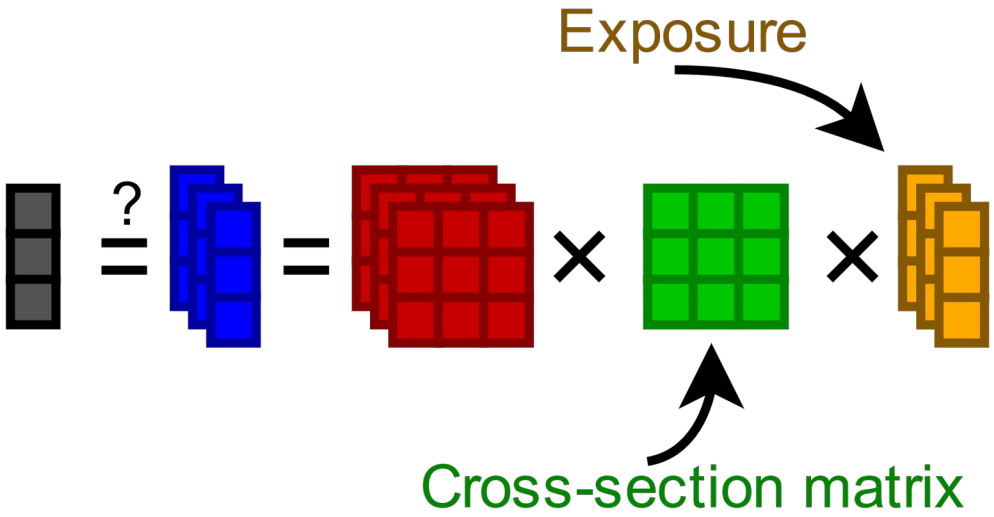


- Background without a reasonable truth definition can be added as pure reco shape in the matrix
- Background strength is single bin in truth vector
- Equivalent to using background template for physics-like background
  - Not possible to change BG model in the future

From events to cross sections

- So far implicitly only talked about event rate predictions
- Models predict cross sections
- Need flux and target mass to go from cross section to event rates
- Problem: We do not know those numbers
  - Flux uncertainty
  - Fiducial mass uncertainty
- Cannot compare prediction using nominal flux with data “using” real flux!
- Give data users tools to vary the flux prediction

Flux forward folding



- Model predicts cross section for each flux bin
- Provide set of flux exposures according to uncertainties
  - Exposure = flux time target mass
- Flux and detector uncertainties can be correlated
  - Make one response matrix correspond to one exposure vector

Response Matrix Utilities

- Implements all of this
- Input:
  - Toy variations of selection (detector systematics)
  - Truth and reco binning
- Provides methods to:
  - Bin data in very flexible binning classes
  - Build matrix and evaluate uncertainty on elements
  - Forward-fold models to reco space
  - Compare to data (e.g. compute likelihoods, p-values, MCMC)
- Pure python (+ standard scientific packages numpy, etc)
  - Easy to install and use
  - $ pip install remu
- Tell me what you expect/want/need!

Further reading

- ReMU documentation: [remu.readthedocs.io]
  - Short introduction to forward folding
  - Examples showing how to use the software
  - Documentation of all classes and functions
- Forward folding method paper: [arXiv:1903.06568]
  - In depth description of the concept and all the maths
- Talk to me [lukas.koch@uni-mainz.de]

Thank you!