STATISTICAL METHODS FOR CROSS-SECTION MEASUREMENTS: PAST, PRESENT AND FUTURE

Lukas Koch
NuSTEC 2024 Summer School

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

DISCLAIMER: OPINIONS!

- Necessarily more familiar with some methods compared to others
  - Biased sample of methods previously/currently in use
  - **Very** biased sample of potential future developments
- If anything seems fishy, probably my fault
- Many subtleties at every step
  - I do not have the time to get into
- Open to Bayesian methods, but biased towards Frequentism
- Most probable answer in statistics: “It depends”

SOME NOTATION

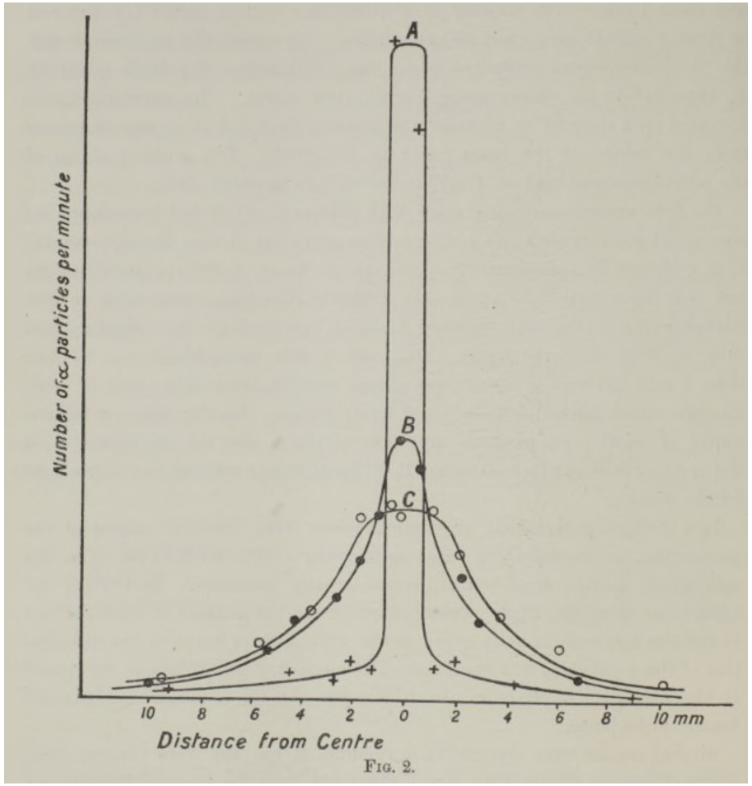
- $\nu_i = \Sigma_j R_{ij} \mu_j$
  - Expected number of observed events ν_i in reco bin i
  - Expected number of true events μ_j in truth bin j
  - Response matrix R is N x M matrix
- Observed events: $n_i \sim \text{Poisson}(\nu_i)$
- True events: $m_j \sim \text{Poisson}(\mu_j)$
- Binned in multiple variables
- Not necessarily same physical meaning
  - track_length_reco = R * momentum_true
- Purely mathematical approach: R = P(event in reco i | event in truth j) = S * eff
- Background handling approaches
  - Subtract from observed events: $n_i = o_i – b_i$
    - “Breaks” Poisson statistics
  - Add to expectation $\nu_i = \varsigma_i + \beta_i$

EVENT RATES VS CROSS SECTIONS

- $\mu_j = \Sigma_k\, T\, (d\sigma/dy)_{jk}\, \Phi_k\, \Delta y_j = T\, (d\sigma/dy)_{j,\Phi\text{-avg}}\, \Phi\, \Delta y_j$
  - For “thin” targets
    - For a neutrino, “thin” can mean a lightyear of lead
  - Assuming cross section is sufficiently constant over bin!
- Conceptual steps:
  - Measure n_i → Use it as proxy for ν_i
  - Unfold and efficiency correct to μ_j
  - Convert event rates to cross sections
- Uncertainties break neat factorisation
  - E.g. detector smearing depends on neutrino flux uncertainty?
- Details vary a lot: “It depends”

JUST LOOK AT RECO

- Implicitly compare n_i with μ_j
  - Pretend y_{reco} and y_{truth} are the same
- Ancient past: Don't even put error bars
  - Not as unreasonable as it sounds
    - **n** vs. **ν**
- Slight improvement: bin-by-bin efficiency correction: n_i / eff_i
  - Only does what you expect if R is diagonal → No smearing



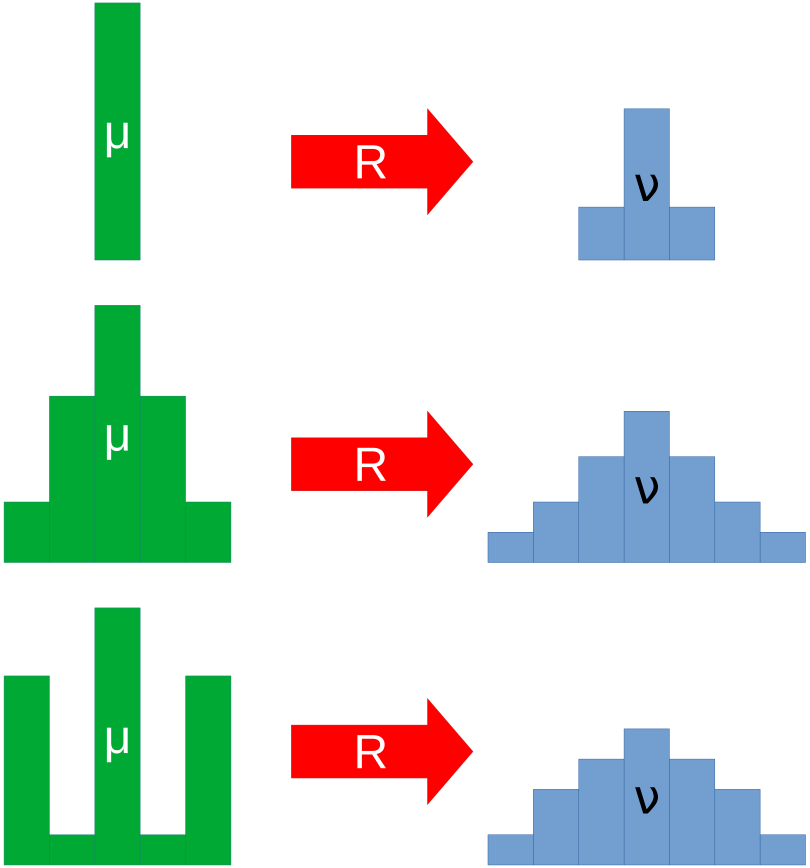
H. Geiger, On the scattering of the α-particles by matter, https://doi.org/10.1098/rspa.1908.0067

NAIVE APROACH: JUST INVERT R

- Usually we have smearing
- $\boldsymbol{\nu} = R\boldsymbol{\mu}$ so why not just calculate $\boldsymbol{\mu} = R^{-1}\,\boldsymbol{\nu} \approx R^{-1}\,\mathbf{n}$
- Possible when $N = M$
  - Choose suitable left-inverse when $N > M$
- Solves least squares problem:
  - Minimize $|\boldsymbol{\nu} - \mathbf{n}|^2 = |R\boldsymbol{\mu} - \mathbf{n}|^2$
  - $\hat{\boldsymbol{\mu}} = (R^T R)^{-1} R^T\,\mathbf{n} = R^{-1}\,\mathbf{n}$
  - Equivalent to maximum likelihood solution when uncertainties Gaussian with known variances
- Can lead to large variance and strong anticorrelations in result

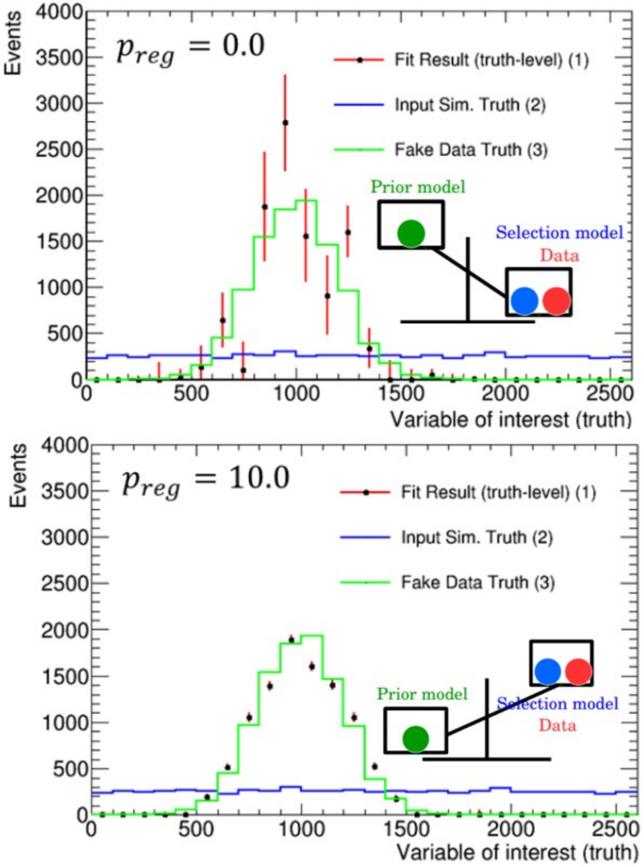
THE ILL POSED PROBLEM

- Strong correlations stem from fact that very different **μ** can lead to very similar **ν**
- Small fluctuations in **n** lead to large swings in “best guess” at **μ**
- Many different solutions are virtually indistinguishable
  - Pick a nicer looking one!
- Impose a slight preference for “nice looking” results
  - Can be interpreted as Bayesian prior or Frequentist external constraint



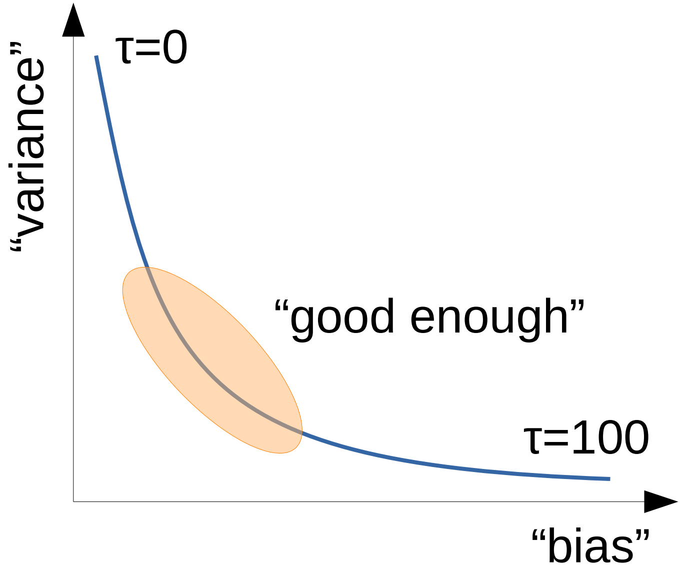
RIDGE REGRESSION / TIKHONOV REGULARISATION

- Modify optimisation problem
  - Add a penalty term for “bad looking” solutions
  - Minimize $|R\boldsymbol{\mu} - \mathbf{n}|^2 + \underline{|C\boldsymbol{\mu}|^2}$
  - $|C\boldsymbol{\mu}|^2 = \boldsymbol{\mu}^T C^T C \boldsymbol{\mu} = \boldsymbol{\mu}^T Q \boldsymbol{\mu}$
- Tikhonov matrix C, or penalty matrix Q
  - Notations vary
  - Choice of C/Q determines what is penalised and how strongly, e.g.
    - $Q = \tau\,\mathbf{I}$ → L_2 norm of $\boldsymbol{\mu}$
    - $\boldsymbol{\mu}^T Q \boldsymbol{\mu} = \tau\,\Sigma(\mu_j - \mu_{(j+1)})^2$
      → Squared differences of neighbouring bins
- New solution
  - $\hat{\boldsymbol{\mu}} = (R^T R + Q)^{-1} R^T\,\mathbf{n}$
  - Adding Q makes $R^T R$ “less problematic” to invert



Borrowed from S. Dolan

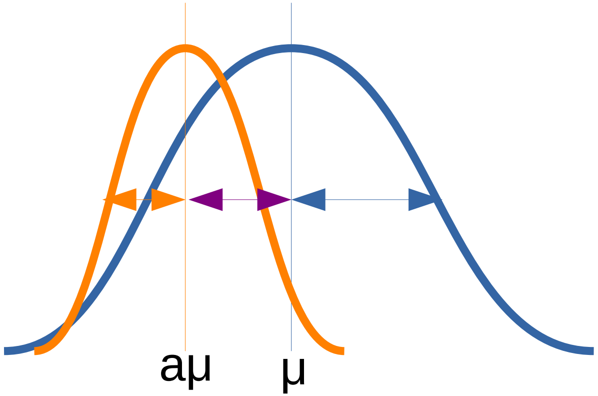
HOW STRONGLY TO REGULARISE



- Regularisation can be seen as prior/external constraint
  - Should be well defined
- Mostly it is introduced ad-hoc
  - Might know what we dislike, but not how much
  - Regularisation strength τ not known a priori
- Regularisation introduces bias
  - Also messes with coverage properties
- Usually some heuristic method to “balance” bias and variance of result
  - e.g. L-curve method
- Can define an objective function and optimize with respect to it
  - What should be optimized can be subjective

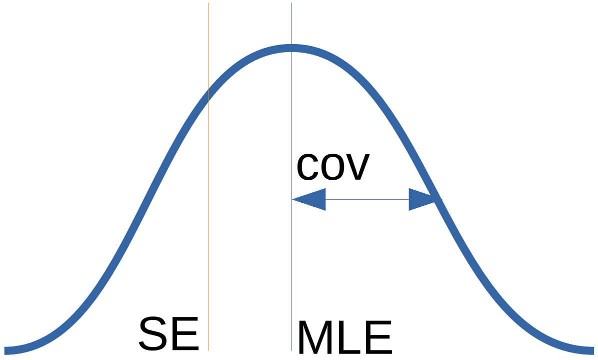
STATISTICAL SHRINKAGE

- Why is it reasonable to penalise large $|\boldsymbol{\mu}|^2$?
- E.g. want to estimate mean value of normal distirbution
- Single sample x from N(μ, σ)
  - Maximum likelihood estimator (MLE): $\hat{\mu} = x$
  - $E[(x-\mu)^2] = \sigma^2$
- Multiply x by shrinkage factor a
  - Shrinkage esitmator (SE): $\hat{\mu} = ax$
  - $E[(ax-\mu)^2] = (a-1)^2\mu^2 + a^2\sigma^2$
  - Minimal at $a = \mu^2 / (\sigma^2 + \mu^2) < 1$
- SE reduces expected squared deviation from true mean compared to MLE!
  - At cost of biasing point estimate towards 0
- Choosing a point estimator does not affect the likelihood function



POINT ESTIMATE VS LIKELIHOOD FUNCTION

- But all information of experiment is (should be) inside likelihood function
  - Often approximated as MLE and covariance matrix
  - It is what it is, even if we do not like how it looks
- Understand regularisation as shrinkage
  - Picking a “reasonable” **point estimate**
  - **Not** to regularise the likelihood function
- Regularised covariance just a visualisation tool?
  - Pick a subset of the allowed region around the point estimate
  - Less correlations, less confusing plots
- Need both for full picture
  - Unregularised data release for “undiluted” likelihood function
  - Regularised result as “better” point estimate
  - Consensus for long time that it would be good to publish likelihood functions
    - Used both in Bayesian and Frequentist analyses



WIENER SVD

https://doi.org/10.1088/1748-0221/12/10/P10002

- Singular Value Decompostion (SVD) can be used to get left inverse of R and solve the least squares problem
- Apply Wiener filter which maximises signal to noise ratio
  - Assuming a given signal shape
  - Inspired by signal processing
  - This is the regularisation
- No tunable regularisation strength
  - Already “optimized” for the signal to noise ratio

RELATION TO UNREGULARISED RESULT

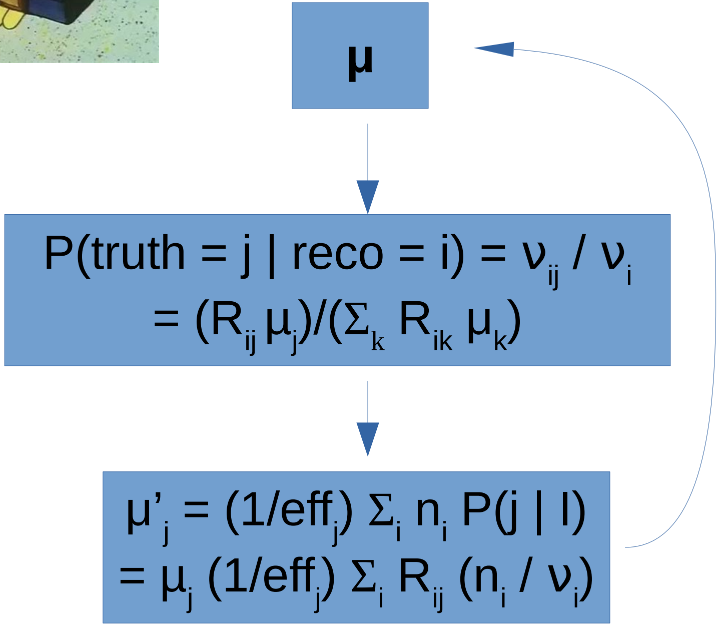
- Wiener SVD yields “additional smearing matrix” A
- It relates regularised result to unregularised one
  - $\boldsymbol{\mu}' = A\,\boldsymbol{\mu}$
    - Does this remind you of the shrinkage estimator?
  - $V' = AVA^T$
- No need to privde two separate results!
  - Just publish A together with either ($\boldsymbol{\mu}$, V) or ($\boldsymbol{\mu}'$, V’)
- Better call A “regularisation matrix”?
  - Does not conserve event numbers and can have negative elements

ITERATIVE UNFOLDING / D'AGOSTINI METHOD

https://arxiv.org/abs/1010.0632

- Also known as Bayesian unfolding
  - Should we be calling it that?
  - It is Bayesian update of priors for 1 iteration
  - It approaches matrix inversion result for inf iterations
    (as long as all $\hat{\mu}$ are positive)
  - "Squeezing the data multiple times" for everything in between?
- # of iterations determines regularisation!
  - Low # →"remembers" first prior → strong regularisation
  - (# → inf) → "forgets" first prior → no regularisation
    - Assuming no smoothing in between iterations

Simplified:



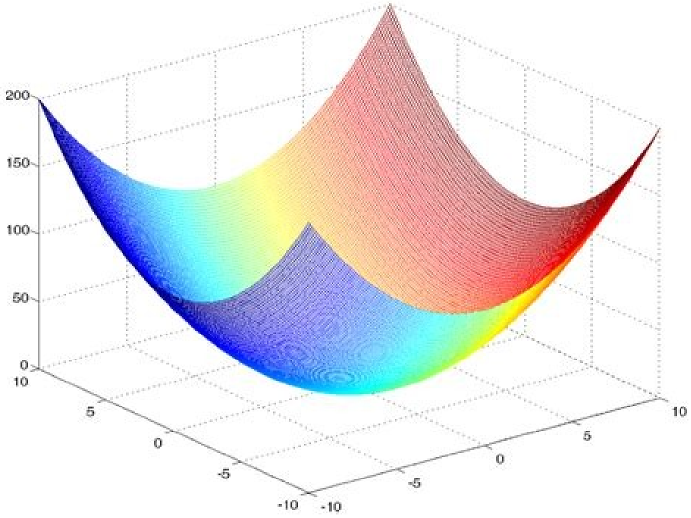
μ

$P(\text{truth} = j \mid \text{reco} = i) = \nu_{ij} / \nu_i$
$= (R_{ij}\, \mu_j)/(\Sigma_k\, R_{ik}\, \mu_k)$

$\mu'_j = (1/\text{eff}_j)\, \Sigma_i\, n_i\, P(j \mid I)$
$= \mu_j\, (1/\text{eff}_j)\, \Sigma_i\, R_{ij}\, (n_i / \nu_i)$

LIKELIHOOD FITTING

E.g. https://arxiv.org/abs/2303.14228



- Explicitly treat problem as parameter fit
  - Poisson likelihood in reco bins
  - Parameters of interest **θ** that scale cross section in truth bins
  - Systematic nuisance parameters **φ**
    - Constrained by “priors” = external constraints
  - “Just” need a function -2 log L(**θ**, **φ** | **n**) and a minimizer
  - Get MLE & parabolic approximation (covariance)
- Add regularisation / penalty terms explicitly

FREQUENTIST FIT, BAYESIAN PROPAGATION?

- Result of fit contains many nuisance parameters
- Correlated uncertainties need to be propagated to XSECs
- Ideal Frequentist approach
  - For each M-dimensional XSEC, maximise likelihood over parameters
    - Profile likelihood
  - Not trivial
- Pragmatic aproach
  - Throw parameters according to MLE & covariance
  - Calculate XSEC for each throw
  - Usually calculate central value and covariance from sample
    - Could also publish throws in case of non-Gaußian results

ADD REGULARISATION AFTER THE FACT?

https://doi.org/10.1088/1748-0221/17/10/P10021

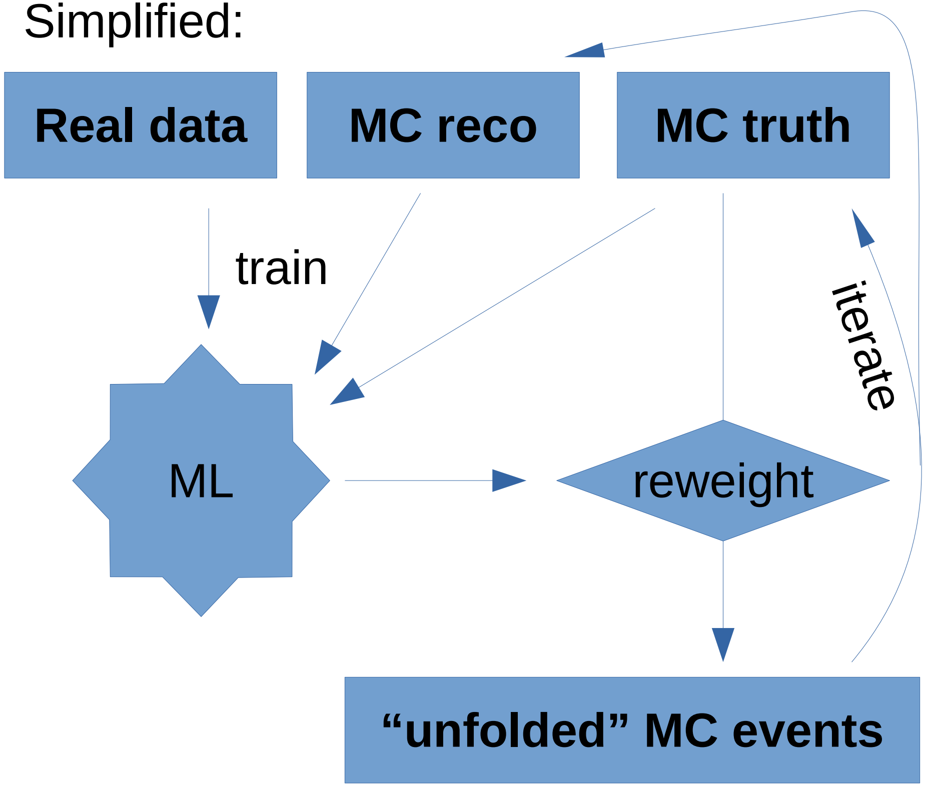
- Take inspiration from Wiener SVD
  - Apply regularisation as a matrix multiplication to the unregularised result
- Given any likelihood described as MLE & covariance, adding a Thikonov penalty term leads to a new result
- Can be applied to any unregularised result → post hoc
  - As long as regularised result is close to unregularised one
    - Parabola approximation of log likelihood stays valid

$$\begin{aligned}
-2\ln(L(\theta)) &\approx (\theta-\hat{\theta})^T V^{-1}(\theta-\hat{\theta}) + const.\\
P(\theta) &= \theta^T Q\,\theta\\
-2\ln(L'(\theta)) &= -2\ln(L(\theta)) + P(\theta)\\
&\approx (\theta-\hat{\theta}')^T V'^{-1}(\theta-\hat{\theta}') + const.\\
\hat{\theta}' &= A\,\hat{\theta}\\
V' &= A\,V\,A^T\\
A &= (V^{-1}+Q)^{-1}V^{-1}
\end{aligned}$$

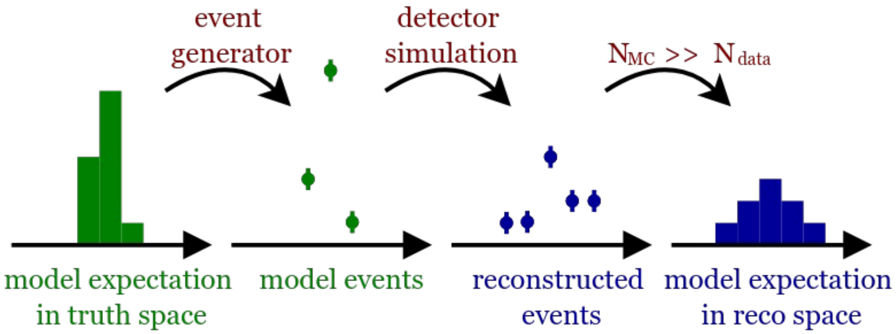
OMNIFOLD (AS I UNDERSTAND IT)

- Use Machine Learning (ML) techniques to create MC reweighter to match MC to measured reco data
  - Based on un-binned event properties
- Re-weighted MC is the “unfolded” result!
  - Can be binneed in any way desired to report a XSEC
- Cutting edge research
  - Just about ready for production use?
  - We will hear more this week!

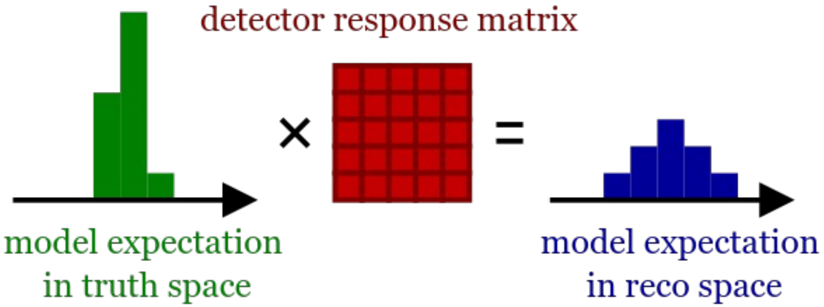
Simplified:



BACK TO THE ROOTS

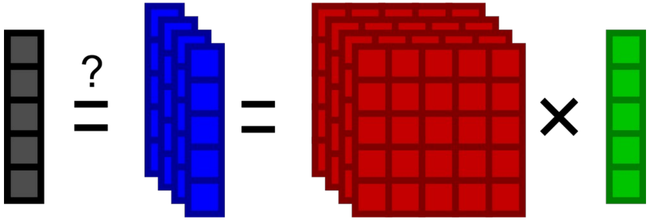


- Possible to do science without unfolding
- Compare models with data in reco space
  - But consider detector effects: Forward folding
  - Allows full statistical analysis
  - The data is exactly what we saw: **n** is aperfectly known fixed number
  - Test whether models are compatible, i.e the predicted **ν**
- How to facilitate use of data by external consumers?
  - Not experts on the detector response
  - No access to (often complicated) simulation frameworks
  - Data needs low entry barrier to be used by many people



SOFWARE AVAILABLE

- Delphes
  - https://cp3.irmp.ucl.ac.be/projects/delphes
  - Developed for collider experiments
- Rivet
  - https://rivet.hepforge.org/
  - Developed for collider experiments
- ReMU – Response Matrix Utilities
  - https://remu.readthedocs.io
  - Developed for neutrino interaction measurements
  - Builds response matrices and uncertainties from MC
  - Fully developed statistical model of detector, flux, and MC stat uncertainties https://iopscience.iop.org/article/10.1088/1748-0221/14/09/P09013



DISCUSSION STARTERS

- Unregularised result is best approximation of Likelihood
  - e.g. for fits and statistical tests of models
- Regularisation should be used to pick a representative point estimate
  - e.g. for plots
- We should always make likelihood function available
  - Unregularised result or something more complicated
  - Wiener SVD and post-hoc regularisation make this trivially easy
    - Added bonus: regularisd and unregularised result are directly related

- Include as many method details as possible in your papers
  - Lots of nuances, caveats, assumptions…
  - Not practical to spell out every single check/study/approximation
    - Or ist it?
    - Dedicated method paper?
  - Have to take papers at face value
    - Trust in what is written
    - Assume the worst about what is not written?
    - Assume the best?
    - Hope for the best but expect the worst?

UNFOLD SOME THINGS!

- Poisson counting experiment w/ 3 bins
- Unfold and estimate true expectation values (MLE & COV)
- Calculate p-values of μ_1 and μ_2
- Forward-fold $\mu_{1/2}$
- Calcualte p-values in reco space
- Play around!
  - In-/decrease smearing
  - Plot results
  - Regularize ...

$$R = \begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0.3 & 0.6 & 0.3 \\ 0.2 & 0.2 & 0.5 \end{pmatrix}$$

$$\mu_1 = \begin{pmatrix} 50 \\ 100 \\ 50 \end{pmatrix}$$

$$n = \begin{pmatrix} 48 \\ 81 \\ 62 \end{pmatrix}$$

$$\mu_2 = \begin{pmatrix} 40 \\ 90 \\ 100 \end{pmatrix}$$

> “Good physicists do have priors and always use them!
> (Only the perfect idiot has no priors.)
>
> – G. D’Agostini
> arXiv:1010.0632

> “Note that venerable proverb:
> Children and fools always speak the truth.
>
> – Mark Twain
> On the Decay of the Art of Lying

Thanks!

Backup

EXAMPLE PENALTY MATRICES

Penalise bin-to-bin differences

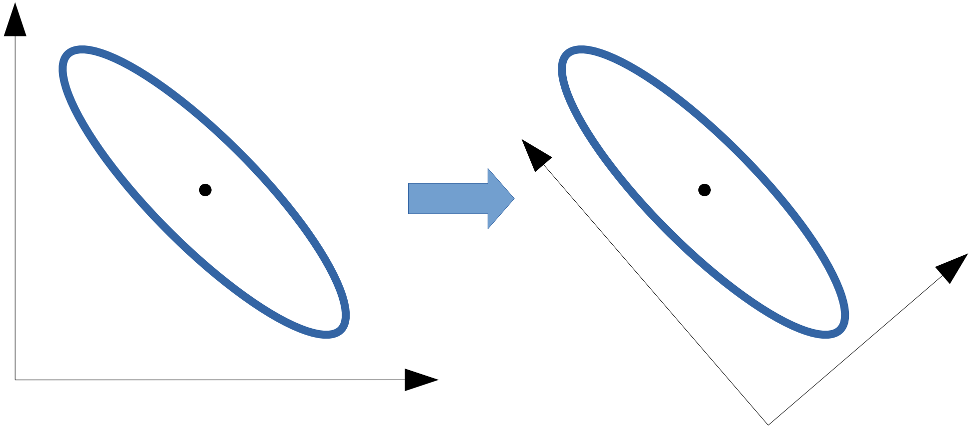
$$\tau Q_1 = \tau \begin{pmatrix} 1 & -1 & 0 & 0 & \\ -1 & 2 & -1 & 0 & \dots \\ 0 & -1 & 2 & -1 & \\ 0 & 0 & -1 & 2 & \\ & \vdots & & & \ddots \end{pmatrix}$$

Penalise bin-to-bin model scaling differences

$$\tau Q_{1m} = \tau \begin{pmatrix} 1/m_1^2 & -1/(m_1 m_2) & 0 & 0 & \\ -1/(m_1 m_2) & 2/m_2^2 & -1/(m_2 m_3) & 0 & \dots \\ 0 & -1/(m_2 m_3) & 2/m_3^2 & -1/(m_3 m_4) & \\ 0 & 0 & -1/(m_3 m_4) & 2/m_4^2 & \\ & \vdots & & & \ddots \end{pmatrix}$$

TWO WAYS OF INTERPRETING A

- Coordinate transformation
- New result describes exactly the same distribution, but with different axes
  - No information lost
- Intuitive in 2D
- Axes of histograms no longer make sense



- Modification of result
- Coordinate axes stay the same, but distribution changes
  - Change of result
- Axes and bin values retain same meaning