Discussions on unfolding problems, methods and solutions – a physics analysis perspective

Bogdan Malaescu
(LPNE Paris - CNRS, CERN)

bogdan.malaescu@cern.ch

17/05/2018
Content

• Introduction: “What is the (un)folding about”
• Choice of phase-space
• Unfolding methods
• Methods for parameter setting, uncertainty evaluation and propagation
• How do we publish our data: HEPData entries for unfolded results
• Physics studies using unfolded data
• Multidimensional unfolding + Choice of binning

ATLAS (internal) recommendations:
https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/StandardModelUnfoldingNew
https://cds.cern.ch/record/1694351
Examples of (un)folding problems: PET scan

Positron Emission Tomography

Y. Vardi et al.
http://www.jstor.org/stable/2288030
Examples of (un)folding problems: top pairs @ LHC

F. Spano

EPJ Web of Conferences 55, 03002 (2013)
Examples of (un)folding problems: jets @ LHC

Jets: “sprays” of (quasi-)stable particles called hadrons, observed in the ATLAS detector → proxy to fundamental interactions in Nature: Explore the limits of SM
Environment and strategy of the jet studies

Typical proton-proton collision: a complex process in a difficult environment

Pile-up

A. Kusina

Beam of partons
- Radiation from incoming partons
- Primary hard scatter
- Radiation from outgoing partons
- Hadronization
- Multiple Inter. / Underlying event

NP corrections
- Hadronization & UE

Calibration + Unfolding
- Jet energy response & resolution

Goal: publish data corrected for detector effects (with minimal bias and minimal model dependence), with the full information needed for comparisons with theory predictions

B. Malaescu – Discussions on unfolding problems, methods and solutions
Detector effects, folding and unfolding

- **Folding**: \( f_{\text{meas}}(x) = \int R(x|y) f_{\text{true}}(y) \, dy \)

\[
P_{i,j} = \frac{A_{i,j}}{\sum_{k=1}^{n_d} A_{k,j}} \quad ; \quad d = P \cdot t
\]

- Focus on unfolding of detector effects (acceptance correction factorized)
- Unfolding is not a simple numerical problem
  \rightarrow **regularization methods** necessary
Choice of the phase-space

- Selection defining phase-space at “truth” level – as close as possible to the reconstructed-level selection: *minimize extrapolation to reduce model dependence*

- Include over-/under-flow bins when migrations to the region of interest are relevant → These extra bins are generally not published

arXiv:1112.6297
Unfolding methods + choice of one method & settings

- **Maximum likelihood / matrix inversion**
- **SVD ( + Tikhonov regularization )**
- **Iterative Bayes-inspired regularized unfolding**
- **Full Bayesian unfolding**
- **Iterative, dynamically stabilized (IDS) method**

**Bin-by-bin correction**: \( d_i \cdot (T_i / R_i)^{MC} \) → potentially large bias by relying on truth MC (used only when small bin-to-bin migrations & for statistics limited measurements e.g. Higgs differential Xsec; cross-check with matrix-based method)

- In general, recommended not to (dis)favor some particular method
- Recommended to evaluate the performance of several methods & regularizations and use the “optimal” one for the given use-case

→ Take into account: systematic uncertainty related to the unfolding method (bias due to MC/data shape difference & regularization); impact on statistical uncertainties & correlations; constraints induced on binning choice
Matrix inversion

Folding of signal and background in data:

\[ E[n] = \nu = R\mu + \beta \]

Unfolding based on matrix inversion:

\[ \hat{\mu} = R^{-1}(n - \beta) \]

→ Result unbiased, but with large variances
Singular Value Decomposition (SVD)

→ Inspired by the matrix inversion, but with regularization:
Suppress effect of small eigenvalues (~noise) + constraint on smoothness of the unfolded distribution → Regularization (may introduce bias)

\[ S(\boldsymbol{\mu}) = - \sum_{i=1}^{M-2} \left[ (\mu_{i+2} - \mu_{i+1}) - (\mu_{i+1} - \mu_i) \right]^2 \]

Singular Value Decomposition (SVD)

→ Inspired by the matrix inversion, but with regularization:
Suppress effect of small eigenvalues (∼noise) + constraint on smoothness of the
unfolded distribution → Regularization (may introduce bias)

\[ S(\mu) = -\sum_{i=1}^{M-2} [(\mu_{i+2} - \mu_{i+1}) - (\mu_{i+1} - \mu_i)]^2 \]


B. Malaescu – Discussions on unfolding problems, methods and solutions
An Iterative, Bayes-inspired Unfolding Method

\[ P_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n_d} A_{kj}} \]

\[ \tilde{P}_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n_u} A_{ik}} ; u = \tilde{P} \cdot d \]

→ Note: \( \tilde{P}_{ij} \) depends on the shape of the truth distribution in MC

• 1st unfolding, where the original transfer matrix is used

→ 1) Transfer matrix improvement (hence of the unfolding probability matrix)

Reweight the truth MC distribution based on previous unfolding result.

→ Other methods exist, like e.g. dynamical regularization from the treatment of fluctuations in each bin, at each step of the procedure
Iterative methods: choice of the number of iterations

- Number of iterations = regularization parameter (the corresponding recommendations apply)

- Compare data and the modified reconstructed MC: see how much information is left to be propagated from the data shape to the truth MC shape → bin-by-bin comparison or using a $\chi^2$

- Suggestion by d'Agostini: compare results from consecutive steps → risk of ~small changes between consecutive steps, while having a significant bias
Statistical uncertainties

- Due to both data and MC
- Propagated using pseudo-experiments done separately/simultaneously for data and MC

→ Bootstrap method
  - multiply event weights
    - by random number: Poisson(1)
    - seed given by event number
  - allows to correlate measurements with overlapping samples

- Publish covariance matrix and/or a series of results based on each pseudo-experiment (i.e. Bootstrap replicas)

- Some unfolding methods provide estimates of the stat uncertainties

→ recommend cross-check with pseudo-experiments
Propagation of systematic uncertainties from inputs

- Modify input (pseudo-)data spectrum by $\pm 1\sigma$ of the uncertainty, re-do unfolding and compare with nominal result
  
  → can also use $1...5\sigma$ scans or pseudo-experiments

- Can shift reconstructed spectrum in transfer matrix instead of input spectrum: switched positive and negative variations

- For resolution uncertainties, perform smearing of the transfer matrix: smearing factor given by quadratic difference between resolution enhanced by $1\sigma$ and nominal resolution
Propagation of systematic uncertainties from inputs

• Bootstrap method to evaluate statistical uncertainties on the propagated systematics + rebinning/smoothing; [arXiv:1312.3524]

• Alternative propagation using pseudo-experiments (more difficult to probe e.g. 5σ effects)

• Alternative propagation option: include uncertainties as nuisance parameters in the definition of the response matrix + profile likelihood or Bayesian marginalization (often used for folding/template fits)
Propagation of systematic uncertainties from inputs

- Split of systematics in sub-components (fully correlated in phase-space, independent between each-other) allows to evaluate correlations between different phase-space regions and between different measurements

- Information made available in HEPData

\[ Cov_{ij} = \sum_{k=1}^{N_{syst}} s_i^k \cdot s_j^k \]

\[ \sqrt{s}=13 \text{ TeV, } 81 \text{ nb}^{-1} - 3.2 \text{ fb}^{-1} \]

\[ \text{ATLAS} \]

\[ \text{Dijets} \quad \text{Inclusive jets} \]

\[ \text{Correlation total uncertainty} \]

arXiv:1711.02692
Tests of the unfolding

• “Technical closure test” → same MC for the transfer matrix and input distribution (pseudo-data)

• “Data-driven closure test” → allows to evaluate a systematic related to the unfolding method and the choice of regularization (see next slides)

• “Linearity test” → MC samples with various truth inputs; check linear dependence between unfolded and truth values of a quantity of interest

• “Pull test” → relevant only for unfolding methods providing an estimate of the stat uncertainty (i.e. not from pseudo-experiments)
Data-driven closure test\(^*\): motivation, procedure, example

- In-situ determination of the unfolding uncertainty related to the MC shape (i.e. to the data/MC shape difference) and to the regularization:
  - reweight true MC by smooth function: improved data/recoMC agreement;
  - unfold the reweighted reconstructed MC;
  - compare with reweighted true MC.

Data-driven closure test(*): motivation, procedure, example

- In-situ determination of the **unfolding uncertainty** related to the MC shape (i.e. to the data/MC shape difference) and to the **regularization**:  
  - reweight true MC by smooth function: improved data/recoMC agreement;
  - unfold the reweighted reconstructed MC;
  - compare with reweighted true MC.

Some subtleties
Data-driven closure test: remarks

- In general, when the data/MC shape differences increase, the unfolding bias increases.

Still, in general, the unfolding bias $<<$ shape differences (see previous slides).

- Folding of modified truth MC can be done with nominal/stat fluctuated resolution matrix: the 2nd option requires a series of toys – median converges towards nominal.

- Data/MC shape comparison: is some shape difference “worth” being taken into account in the reweighting?

  → in general, binned data/MC comparison taking into account stat uncertainties.

  → one can perform more quantitative comparisons (e.g. through $\chi^2$ evaluations) taking into account data and MC systematics.

    - shape comparison: the MC normalization free parameter in the $\chi^2$ evaluations.

    - has to be done globally and more locally (for restricted phase-space regions): in global comparison, local systematic differences can be “hidden” by the ndof++.
Analyses with several “correction steps”

- Analyses generally involve a series of corrections for: efficiency, acceptance, migrations, subtraction of background / "fakes" etc.

- Guidelines:
  → coherent selection criteria applied to the data and the reconstructed MC samples (used to derive corrections), at each correction step;
  → coherence between data and the reconstructed MC for what concerns the variables as a function of which the corrections are derived / applied;
  → coherence between the motivation / derivation / application of the corrections;
  → minimize the model dependence for all the corrections.
• **Background subtraction** (data-driven or estimated with MC)

→ (Generally) **performed before unfolding**, because resolution is generally different for signal and background;

(otherwise, special care is required when building the transfer matrix)

• **Propagation of the systematic uncertainties** (for the calibration(s) these often concern data-MC differences) from the input objects to the final analysis results: *avoid double-counting and/or fake cancellations*

→ propagate systematics either through (pseudo-)data (for input distributions, data-driven background estimates etc.) or MC (transfer matrix, MC-templates etc.), but not both;

→ keep track of the signs when evaluating and propagating the impact of a systematic variation on various inputs.
Possible difficulties: fluctuations due to background subtraction

- A “standard” unfolding could propagate large fluctuations into precise regions of the spectrum
- Can be addressed by taking into account the uncertainties of the data points in the unfolding (used to compute the significance of data-MC differences in each bin - IDS)
Treatment of “hidden variables”

→ Check the impact on the average resolution, when integrating differently over various phase-space regions (e.g. central-forward regions when unfolding pT distribution)

→ Does not concern the closure test, for which this extra reweighting is not needed (probes impact of data/MC shape difference for the observable of interest)

• A first check can be done by comparing unfolding results when using different MC samples (different truth-level distributions of variables other than the unfolded one)

→ truth MC shapes for the variable(s) of interest (i.e. the quantities that are unfolded) must be re-weighted such that they match between different MC samples: avoid double-counting with data-driven closure test

→ if significant differences are seen: compare with relevant resolution uncertainties and identify variables causing the difference

• Re-weight the transfer matrix as a function of more variables than the unfolded one and re-do the unfolding with modified matrix: amount of re-weighting should be data-driven
How we report the results in HEPData
W+jets @ 7TeV (2011) – HEPData information
http://hepdata.cedar.ac.uk/view/ins1319490

Split of systematic uncertainties in sub-components
Bootstrap replicas made public & used in e.g. 1602.01110

<table>
<thead>
<tr>
<th>R</th>
<th>RE</th>
<th>P P -&gt; JET X</th>
<th>7000.0 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.3</td>
<td>4230 ± 0.89% (stat) ± 3.0% (sys,jes0)</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>+0.4% -0.5% (sys,jes1) ± 0.2% (sys,jes2)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>± 0.3% (sys,jes3) ± 1.2% (sys,jes4)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.0</td>
<td>± 1.0% (sys,jes5) ± 0.2% -0.1% (sys,jes6)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.5% -0.6% (sys,jes7) ± 0.1% (sys,jes8)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.6% -0.9% (sys,jes9) ± 0.3% (sys,jes10)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.8% (sys,jes11) ± 0.5% -0.6% (sys,jes12)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.2% (sys,jes13) ± 0.0% (sys,jes14)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes15) ± 0.0% (sys,jes16)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes17) ± 0.0% (sys,jes18)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes19) ± 0.0% (sys,jes20)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.1% (sys,jes21) ± 0.1% -0.0% (sys,jes22)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes23) ± 0.1% (sys,jes24)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.1% (sys,jes25) ± 0.3% (sys,jes26)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.4% (sys,jes27) ± 0.5% -0.6% (sys,jes28)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.4% -0.3% (sys,jes29) ± 0.0% (sys,jes30)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes31) ± 0.0% (sys,jes32)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes33) ± 0.1% (sys,jes34)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes35) ± 0.1% (sys,jes36)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes37) ± 0.0% (sys,jes38)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes39) ± 0.0% (sys,jes40)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes41) ± 0.0% (sys,jes42)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes43) ± 0.0% (sys,jes44)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes45) ± 0.0% (sys,jes46)</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>± 0.0% (sys,jes47) ± 0.0% (sys,jes48)</td>
<td></td>
</tr>
</tbody>
</table>

D2(SIG)/DPT(JET)/DYRAP IN PB*GEV**-1

<table>
<thead>
<tr>
<th>PT(JET) IN GEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>4230 ± 0.89% (stat) ± 3.0% (sys,jes0)</td>
</tr>
<tr>
<td>+0.6% -0.7% (sys,jes1) ± 2.4% (sys,jes2)</td>
</tr>
<tr>
<td>+0.4% -0.5% (sys,jes3) ± 0.2% (sys,jes4)</td>
</tr>
<tr>
<td>± 0.3% (sys,jes5) ± 1.2% (sys,jes6)</td>
</tr>
<tr>
<td>± 1.0% (sys,jes7) ± 0.2% -0.1% (sys,jes8)</td>
</tr>
<tr>
<td>± 0.5% -0.6% (sys,jes9) ± 0.1% (sys,jes10)</td>
</tr>
<tr>
<td>± 0.8% -0.9% (sys,jes11) ± 0.3% (sys,jes12)</td>
</tr>
<tr>
<td>± 0.8% (sys,jes13) ± 0.5% -0.6% (sys,jes14)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes15) ± 0.0% (sys,jes16)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes17) ± 0.0% (sys,jes18)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes19) ± 0.0% (sys,jes20)</td>
</tr>
<tr>
<td>± 0.1% (sys,jes21) ± 0.1% -0.0% (sys,jes22)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes23) ± 0.1% (sys,jes24)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes25) ± 0.1% (sys,jes26)</td>
</tr>
<tr>
<td>± 0.1% (sys,jes27) ± 0.3% (sys,jes28)</td>
</tr>
<tr>
<td>± 0.4% (sys,jes29) ± 0.5% -0.6% (sys,jes30)</td>
</tr>
<tr>
<td>± 0.4% -0.3% (sys,jes31) ± 0.0% (sys,jes32)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes33) ± 0.1% (sys,jes34)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes35) ± 0.1% (sys,jes36)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes37) ± 0.0% (sys,jes38)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes39) ± 0.0% (sys,jes40)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes41) ± 0.0% (sys,jes42)</td>
</tr>
<tr>
<td>± 0.0% (sys,jes43) ± 0.0% (sys,jes44)</td>
</tr>
</tbody>
</table>

B. Malaescu – Discussions on unfolding problems, methods and solutions
How we use the results

- PDF and/or $\alpha_s$ fits
- Quantitative data / theory comparisons
- Limits on New Physics contributions
- Etc.
Extraction of Physics information from measurements

→ Involves use of test statistics exploiting information on uncertainties and their correlations (between various measurement bins)

$$\chi^2 (d; t) = \sum_i \left( \frac{d_i - t_i}{\sigma_i(t_i)} \right)^2$$

$$\chi^2 (d; t) = \sum_{i,j} (d_i - t_i) \cdot [C^{-1}(t)]_{ij} \cdot (d_j - t_j)$$
Quantitative data/theory comparisons

- **Generalized** \( \chi^2(d; t) = \min_{\beta_a} \left\{ \sum_{i,j} \left[ d_i - \left( 1 + \sum_a \beta_a \cdot (e^\pm_a(\beta_a))_i \right) t_i \right] \cdot \left[ C^{-1}_{su}(t) \right]_{ij} \right. 

\left. \cdot \left[ d_j - \left( 1 + \sum_a \beta_a \cdot (e^\pm_a(\beta_a))_j \right) t_j \right] + \sum_a \beta_a^2 \right\}, \)

- accounts for correlations and asymmetries of uncertainties

\[ \text{arXiv:1312.3524} \]

- Using **frequentist method to compute p-value**: 

→ pseudo-experiments from theory prediction, with the full information on the uncertainties: build the generalized \( \chi^2 \) distribution (no assumption needed)

→ observed \( \chi^2 \) from the data/theory comparison
PDF comparisons for dijets

- Comparisons to MSTW2008, NNPDF2.3 and ABM11

- Sensitivity to PDFs: level of agreement strongly depends on the PDF set and phase-space region

→ Valuable experimental inputs to constrain proton PDFs: Published information on cross-sections & uncertainties, with their correlations and asymmetries
Limits on New Physics using unfolded distributions

• Explore BSM physics directly at particle level

Contact Interaction Model (CI)
New force mediated by heavy particle

arXiv:1312.3524

• Full frequentist analysis (CLs), with generalized $\chi^2$ as test statistic
→ accounts for correlations and asymmetries of uncertainties (stat. & syst.)
• Limits similar to the ones obtained by dedicated searches (comparing reco-level data with theory predictions folded with detector effects)
nD Unfolding
When to perform an unfolding in nD?

• Decision to perform an unfolding in nD based on:
  → Observables of interest
  → Resolution for the observables of interest
  → Impact of merging of bins of one observable on the average detector response for another one (see discussion on “hidden variables”)
  → Reduced statistics per bin when increasing the number of dimensions / reducing bin size:
    - having more bins enhances the amount of available information
    - some unfolding methods can be impacted in various ways by effects of low statistics
nD unfolding – resolution effects mainly on 1 observable

- Several 2D studies (e.g. inclusive jets Xsec) where the resolution effects mainly impact one observable (pT), but have little impact on the other one (|y|)

→ 1D unfolding for several pT slices

→ small migrations in |y| accounted for through efficiency corrections
In general, even when resolution effects impact several observables, one can convert a nD unfolding exercise into 1D: bin re-ordering

Care needed for adapting regularization methods

→ straightforward for Bayesian-inspired iterative method, IDS, etc. (methods not directly sensitive to the relative position of the bins in the physical space; sensitive only to the amount of migrations between bins)

→ regularization method for SVD has to be adapted (average curvature directly sensitive to the relative position of the bins in the physical space; examples where the method has been adapted to 2D do exist)