Biases and Pitfalls in Unfolding

Igor Volobouev

Texas Tech University

i.volobouev@ttu.edu

July 27, 2023

IML Working Group meeting on unfolding

Why Unfolding is Difficult

- The typical formulation of the unfolding problem is

$$g(\mathbf{y}) = \int K(\mathbf{y}, \mathbf{x}) \lambda(\mathbf{x}) d\mathbf{x}, \tag{1}$$

  where $g(\mathbf{y})$ is the intensity of the observed Poisson process in the space of "smeared" variables $\mathbf{y}$, and $\lambda(\mathbf{x})$ is the intensity we want to reconstruct in the "true" space $\mathbf{x}$. $K(\mathbf{y}, \mathbf{x})$ is known as the kernel or response function. The formulation does not have to be linear but in HEP problems linearity in λ is usually assumed.
- This formulation is intrinsically *infinite-dimensional*. $\lambda(\mathbf{x})$ is an infinite-dimensional parameter whose dimensions are labelled by $\mathbf{x}$ (think field theory). Naturally, an infinite-dimensional parameter can not be reconstructed from finite amount of data. We must make dimensionality reduction assumptions, *i.e.*, *regularize*.
- I prefer to think that "regularization" is any assumption or device that reduces the number of degrees of freedom in the problem. Then, for example, it becomes obvious that binning is regularization.

A Simple Illustration

- In Eq. 1, assume $K(\mathbf{y}, \mathbf{x}) = \mathcal{N}(y - x, \sigma^2)$, where $\mathcal{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2. 1-d problem with a Gaussian resolution function and 100% efficiency.
- If $g(y)$ was exactly known, the Fourier transform of λ could be obtained from $\lambda(\omega) = g(\omega)/K(\omega)$. Note that, in this case, $K(\omega) = e^{-\sigma^2\omega^2/2}$.
- As $g(y)$ is not known, the closest available approximation is the characteristic function of the empirical Poisson intensity of the observed sample, $\rho_e(y) = \sum_{i=1}^{N} \delta(y - y_i)$. Then $\rho_e(\omega) = \int \rho_e(y) e^{i\omega y} dy = \sum_{i=1}^{N} e^{i\omega y_i}$.
- The ratio $\rho_e(\omega)/K(\omega)$ becomes arbitrarily large as $\omega \to \infty$. The "naive" method of estimating $\lambda(\omega)$ as $\rho_e(\omega)/K(\omega)$ thus fails miserably: the high frequency components of the statistical noise contained in $\rho_e(\omega)$ are multiplied by an arbitrarily large factor so that $\rho_e(\omega)/K(\omega)$ is not even square-integrable.
- "Deconvolution density estimation" is the term to google.

Regularization

- All standard approaches to regularization basically do the same thing: they suppress high frequency components of statistical fluctuations at the cost of simultaneously suppressing similar components of the signal. While such approaches are based on the reasonable assumption that the "true" distributions we are reconstructing should be smooth, there are costs in terms of *bias* and *loss of information*.
- The regularization "strength" is usually chosen to optimize some form of the *bias-variance trade-off*.
- The real problem with regularization is that the risks associated with the bias and loss of information become well-defined only at a later stage, when theoretical models are fitted to unfolded data. In the current practice of HEP data analysis, there is a disconnect between the unfolding stage and the model fitting stage. Construction of an optimal solution at the unfolding stage (akin to the Wiener filter) is thereby precluded.

Approaches Based on Binning

- There is a difference between binning and discretization. Binning assumes integration of all relevant functions over some intervals.
- Binning is regularization! Response function $\rightarrow$ response matrix:

$$K_{ij} = \frac{\int_{\mathbf{y}\in b_i} \int_{\mathbf{x}\in b_j} K(\mathbf{y},\mathbf{x})\lambda(\mathbf{x})d\mathbf{x}d\mathbf{y}}{\int_{\mathbf{x}\in b_j} \lambda(\mathbf{x})d\mathbf{x}} \tag{2}$$

- The cost of binning is twofold: loss of information about signal frequencies beyond the Nyquist frequency of the bins (important for sharp peaks) and, as true $\lambda(\mathbf{x})$ is unknown and has to be guessed beforehand, introduction of the "wide bin bias" into the response function. The fraction of events migrating out of the bin to the left and to the right depends on whether the events themselves are concentrated near the left or the right edge of the bin (important for steeply falling spectra).
- Additional regularization becomes necessary if the condition number of K_{ij} is large and the equation $g_i = \sum_j K_{ij}\lambda_j$ is ill-conditioned.

The Problem of Nuisance Parameters

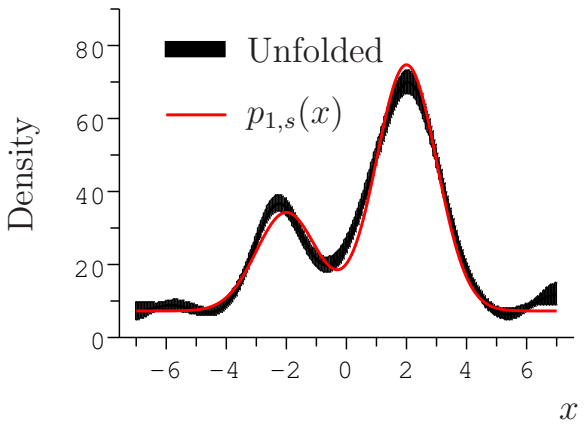
- $K(\mathbf{y}, \mathbf{x})$ is a conditional probability density (with the caveat that it incorporates efficiency): $K(\mathbf{y}, \mathbf{x}) = K(\mathbf{y}|\mathbf{x})$. In real HEP applications, we almost always have $K(\mathbf{y}, \mathbf{x}) = K(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. The character of K dependence on nuisance parameters $\boldsymbol{\theta}$ can be further classified as
  - The calibration problem: elements of $\boldsymbol{\theta}$ are independent from the particle process under study. Not specific to unfolding.
  - The problem of insufficient causation: elements of $\boldsymbol{\theta}$ depend on the process under study. In the absence of direct causation of $\mathbf{y}$ by $\mathbf{x}$, additional assumptions invade the basic formulation of the unfolding problem. Naturally, unfolded results will depend on these assumptions. A simple example: unfolding of jet p_{T} spectra is affected by their assumed η spectra as the jet p_{T} resolution is a function of both p_{T} and η.
- Together with the wide bin bias, the problem of insufficient causation affects a number of LHC analyses which derive their response functions under the Standard Model assumptions and then pretend that the unfolded results could be used to constrain BSM physics.

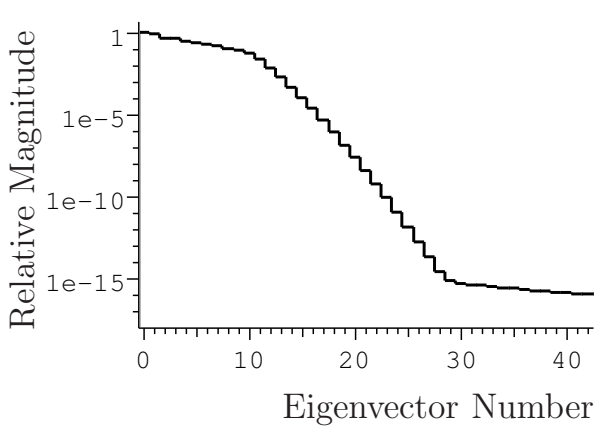
Other Issues

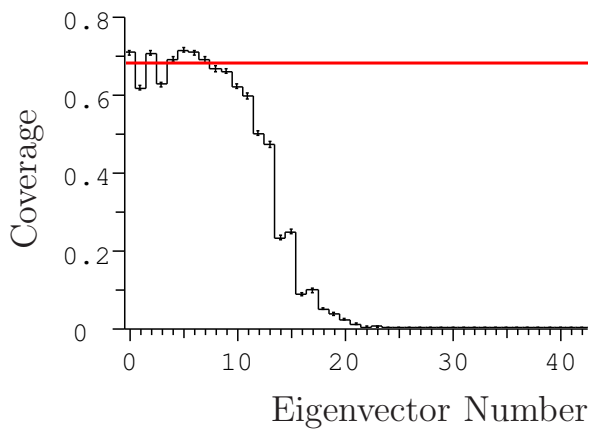
- Statistical uncertainties in the determination of $K(\mathbf{y}, \mathbf{x})$ and K_{ij}.
- Correct handling of the edge effects requires special attention.
- For the χ^2 statistic calculated in binned scenarios, nDoF $\neq$ nBins.
- If the choice of regularization strength is data-driven (as it should be), an additional statistical uncertainty has to be assigned to it. This uncertainty is not taken into account by the error propagation techniques built into the current software.
- A number of techniques incorporate a penalty on deviation from a prior, *e.g.*, implementations of SVD and Richardson-Lucy (a.k.a. D'Agostini) unfolding in "root". Results obtained with these methods can be severely biased towards that prior. As priors are usually obtained by fitting models to previous results, proper combination of the new results with the old ones is no longer possible.
- The null space and the effective null space of the problem are virtually never analyzed. This subject remains unexplored in the context of HEP data analyses.

The Danger of Biased Estimators

- The statistical uncertainties of biased estimators are not subject to the Cramer-Rao bound and do not represent the total error. The statistical covariance matrices determined by linear error propagation are ill-conditioned or singular. The example below comes from arXiv:1408.6500 (Richardson-Lucy unfolding with smoothing).







- In principle, the bias can be accounted for by the appropriate systematic uncertainty. In practice, determination of this uncertainty is difficult and very subjective.
- The degree to which underestimation of total uncertainty is detrimental depends on the subsequent use of the unfolded result.

New Unfolding Techniques

Two basic principles should be kept in mind when new unfolding methods are developed:

- Reasonable frequentist coverage has to be demonstrated.
- The unfolded results have to be presented in such a manner that results obtained by multiple independent measurements can be combined.

Techniques that do not adhere to these principles are not mature enough for LHC data analyses. If your purpose is obtaining physics results (rather than development of statistical and/or ML unfolding methodology), don't waste your time on them.

Unfolding with ML

- A Living Review of Machine Learning for Particle and Nuclear Physics has the "Unfolding" section which currently lists 21 references. Two of them refer to early publications proposing unbinned unfolding techniques that do not rely on ML.
- My own personal quick filter: search the paper pdf for the terms "covariance" and "correlation" referring to matrices as well as for the word "coverage" referring to frequentist coverage. If none of these terms are present in the body of the manuscript, the method that the paper is advocating is probably not ready for the prime time.
- 6 out of 19 papers pass the quick filter. None of them attempts a frequentist coverage study.

Comments on OmniFold

- The original OmniFold paper (see also representative talks by the authors here and here) did not describe how to derive statistical uncertainties of the unfolded results and did not pass the quick filter. Nevertheless, OmniFold was used in this experimental paper by the H1 Collaboration which, of course, had to take care of the statistical uncertainties. They were derived by resampling the data 100 times (MC sample was kept fixed).
- How many samples are needed in order to estimate a covariance matrix reliably? This is actually an interesting question. The answer here states that, for a $n \times n$ matrix, you need at least $25n$ samples. H1, grouping the measured cross-sections into 25 bins, falls short.
- Biases (and the corresponding systematic uncertainties) due to the introduction of regularization remain a mystery. OmniFold estimates the density ratio between the data and the simulation by a classifier DNN. This density ratio estimate is regularized by choosing the network architecture and the stopping rule for classifier training. At least to me, the consequences of this are not at all transparent.

Is There a Better Way to Unfold?

- Note that, when theoretical models are fitted to unfolded results, regularization happens automatically, simply due to the fact that reasonable models have a limited number of parameters. Therefore, the goal of unfolding should be to present results in a manner that minimizes bias and loss of information and avoids regularization as much as possible (this line of thought originally comes from Volker Blobel).
- Since we can't determine $\lambda(\mathbf{x})$ in an unbiased manner, a better approach is to represent the unfolded results by estimating a set of functionals derived from $\lambda(\mathbf{x})$. Ideally, the functionals should be chosen in such a way that their estimates possess the following properties:
  1. Unbiasedness
  2. Proper frequentist coverage
  3. Minimal loss of information by the complete set
  4. Mutual statistical independence (or, at least, absence of correlations)
- I call this the *delayed regularization* approach. More details about it can be found in this talk.

Conclusions

- In unfolding scenarios, care and sofistication are necessary to extract maximum amount of information from your data and to present it in a statistically sound manner. There is no one-size-fits-all recipe.
- While the flexibility and power of ML tools promise substantial improvements, the developed methods must stay firmly grounded in solid statistical principles.
- In classical unfolding techniques, the cornerstone principle of the regularization strength selection is the bias-variance trade-off. It would be very useful to elucidate it in the ML-based unfolding.
- Perhaps, more thought should be given to the minimum regularization/delayed regularization approach.