Iterative Unfolding Optimization with the Mean Squared Error Metric

Shih-Kai Lin

Colorado State University

Finding-numu Meeting
October 1, 2019

Recap

- The standard metric in the ND group used by all analyses requiring unfolding is the average global correlation coefficient[1],

$$\rho_{avg} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{1 - \frac{1}{\boldsymbol{V}_{jj}(\boldsymbol{V}^{-1})_{jj}}} \tag{1}$$

  , where M is the number of bins and $\boldsymbol{V}$ is the covariance matrix in true space inferred by the unfolding algorithm.

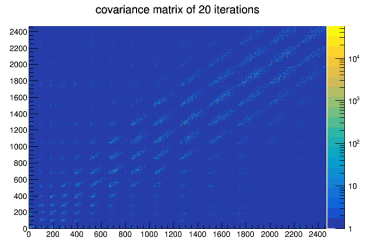
- For analyses with tens of bins, this is a convenient metric. However, for an analysis with thousands of bins, this metric turned out to be infeasible.

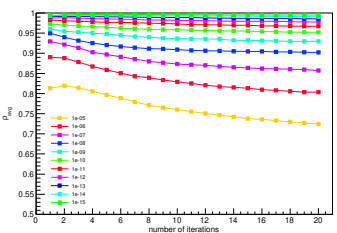
[1] Stefan Schmitt, "Data Unfolding Methods in High Energy Physics"

Infeasibility of Average Global Correlation Coefficient for Many-Bin Analyses

Inverting a covariance matrix this large turns out to be very tricky.

- The covariance matrices all have astronomical condition numbers (i.e., ill-conditioned or nearly singular).
- Numerical inversion is still possible but subject to an arbitrary, small cut-off number, or tolerance, brought into play by SVD.
- Forcefully getting the calculation through results in numerical instability, such as negative values in square root in Eq. 1. Removing unphysical values, results are shown to the right. No clear minimum is observed.



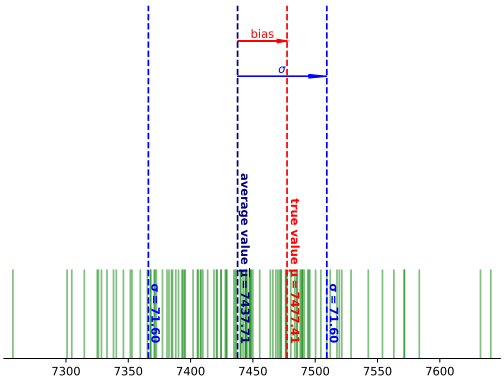


A Simple Alternative Metric: Mean Squared Error

Suppose θ is the a true parameter to be estimated, and $\hat{\theta}$ is an estimator of the parameter. The mean squared error (MSE) can be decomposed into a sum of variance and bias squared.

$$
\begin{aligned}
MSE &= E[(\hat{\theta} - \theta)^2] \\
&= E[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= (E[\hat{\theta}^2] - E[\hat{\theta}]^2) + (E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2) \\
&= V[\hat{\theta}] + b^2 \qquad (2)
\end{aligned}
$$

, where $b = E[\hat{\theta}] - \theta$ is the bias of the estimator. Equation 2 is called bias-variance decomposition. This is a very common metric in statistics and machine learning as well.



Mean Squared Error in the Context of Unfolding

Given a true histogram $\boldsymbol{\mu} = (\mu_1, ..., \mu_M)$, where μ_i's are true counts in the i-th bin, $i = 1, ..., M$, unfolding can be viewed as a procedure that outputs a vector of estimators for the bin counts $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, ..., \hat{\mu}_M)$. Denoting the mean squared error for the i-th bin MSE_i, MSE for the histogram can be defined as

$$\begin{aligned} MSE &= \frac{1}{M}\sum_{i=1}^{M} MSE_i \\ &= \frac{1}{M}\sum_{i=1}^{M} V_{ii} + \hat{b}_i^2 \end{aligned} \tag{3}$$

, where $V_{ii} = cov[\hat{\mu}_i, \hat{\mu}_i]$ and $\hat{b}_i$ is an estimator of $E[\hat{\mu}_i] - \mu_i$.

Very often one wants to estimate more accurately bins with smaller absolute statistical uncertainties. In this case, weighted MSE can be used:

$$\begin{aligned} weighted \quad MSE &= \frac{1}{M}\sum_{i=1}^{M} \frac{MSE_i}{\hat{\mu}_i} \\ &= \frac{1}{M}\sum_{i=1}^{M} \frac{V_{ii} + \hat{b}_i^2}{\hat{\mu}_i} \end{aligned} \tag{4}$$

In this study, a minimum bin count is required to satisfy Poisson statistics, and is compared to the result without count constraint.

Accessing MSE with Toy Monte Carlo

1. Take the true spectrum. Generate 100 pseudo experiments by fluctuating each bin by Poisson with the true bin count as the parameter, i.e., average value.
2. Smear each experiment by applying the normalized migration matrix to the true spectrum.

$$\nu_i = \sum_j \left(\frac{A_{ij}}{\sum_i A_{ij}} \right) \mu_j \tag{5}$$

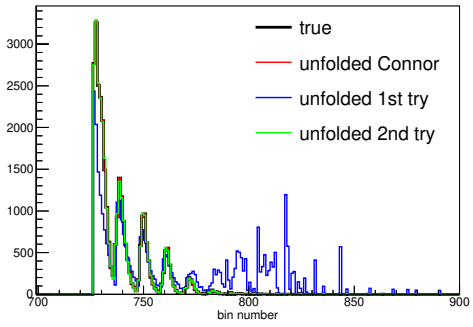
   , where ν_i is the reco spectrum and μ_j is the true spectrum.
3. Unfold each spectrum by a certain number of iterations.
4. For each iteration, calculate MSE with the 100 spectra. Find the number of iterations with a minimum MSE.

Why My First Attempt Did Not Work Out

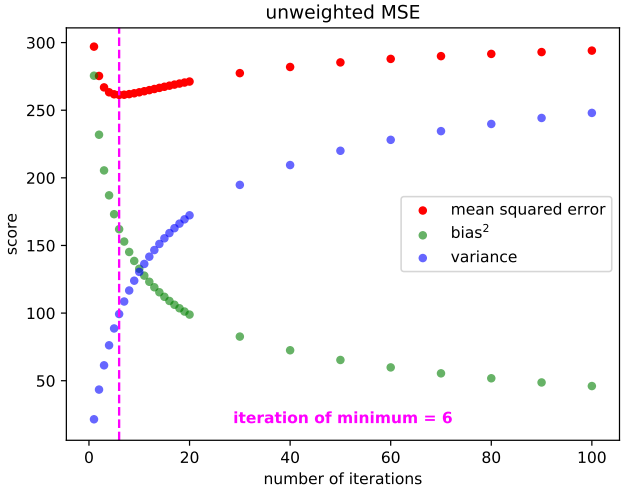
- My first try led to exceedingly large bias, washing out the variance term completely.
- All ND analyzers have been using the "migration matrix" obtained by `CAFAna` directly. My first attempt used the "normalized" migration matrix instead, resulting in large unfolded counts at bins with almost no counts in true space.

$$B_{ij} = \frac{A_{ij}}{\sum_i A_{ij}}$$

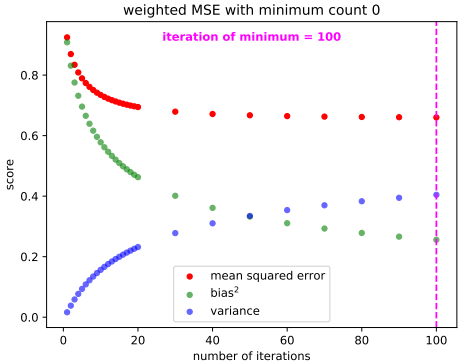
- I used the raw "migration matrix" in my second attempt and got the same results as others. Probably double counting somewhere.

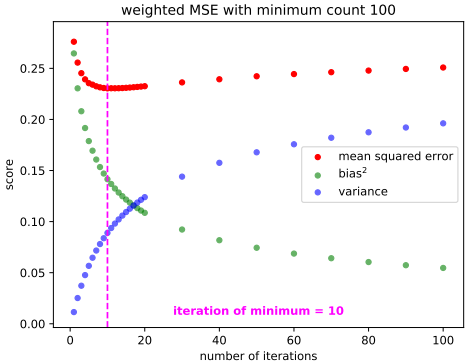


Results: Unweighted MSE



Results: Weighted MSE





Reference

Most of the contents in this document are taken from this textbook, especially Chapter 11 dedicated to unfolding.