Carnegie Mellon University

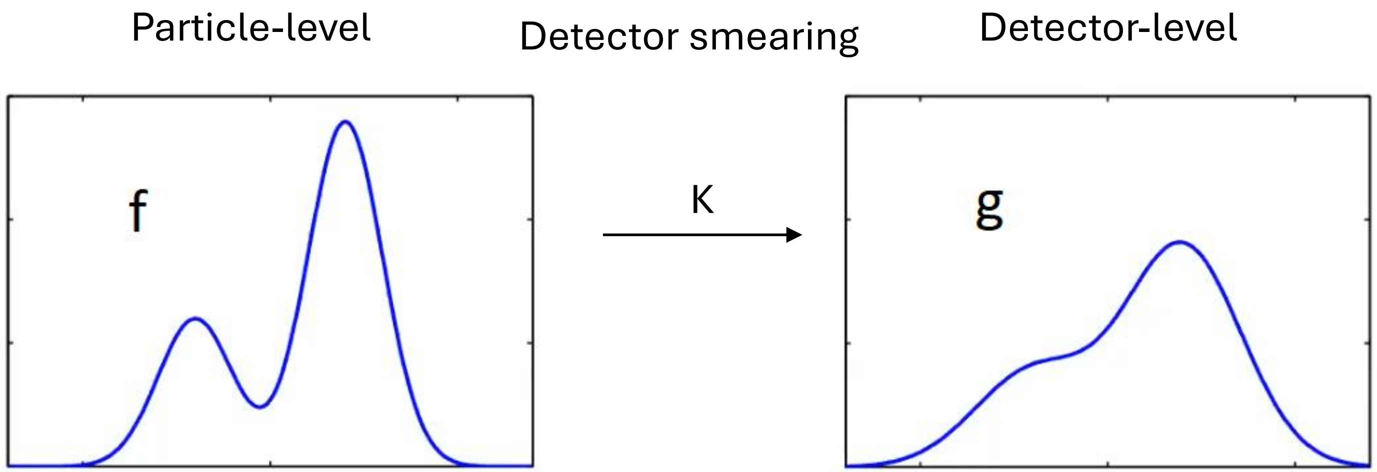
Response Matrix Estimation in Unfolding Differential Cross Sections

JUNE 11, 2024

Richard Zhu
Mikael Kuusela
Larry Wasserman
Department of Statistics and Data Science
Carnegie Mellon University

France-Berkeley PHYSTAT Conference on Unfolding

Forward model for unfolding



$$g(y) = \int_{x \in T} k(y,x) f(x) dx\,, \qquad k(y,x) = p(smeared\ observation\ y\mid true\ event\ x)$$

Discretization

Let $\{T_j\}_{j=1}^{n}$ be a partition of the particle-level space T and $\{S_i\}_{i=1}^{m}$ be a partition of the detector-level space S.

$$f \to \boldsymbol{\lambda}, g \to \boldsymbol{\mu}$$

$$\boldsymbol{\lambda} = \left[\int_{T_1} f(x)dx, \dots, \int_{T_n} f(x)dx\right], \boldsymbol{\mu} = \left[\int_{S_1} g(y)dy, \dots, \int_{S_m} g(y)dy\right]$$

$\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ where the elements of the response matrix $\boldsymbol{K}$ are given by

$$K_{ij} = \frac{\int_{y \in S_i} \int_{x \in T_j} k(y, x) f(x) dx dy}{\int_{y \in S_i} f(x) dx}$$

$$= P(smeared\ observation\ in\ bin\ i \mid true\ event\ in\ bin\ j)$$

Goal: Inference on the true mean $\boldsymbol{\lambda}$

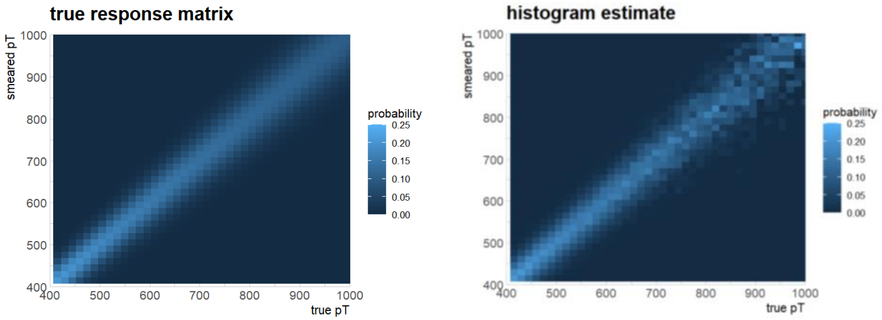
Statistical uncertainty in the response matrix

- The response matrix $\boldsymbol{K}$ is usually not known analytically, but instead estimated with Monte Carlo simulation, which introduces **statistical uncertainty** on $\boldsymbol{K}$.

- Traditionally, this has been estimated by binning the true and smeared events and counting the propagation of events between the bins, i.e.

$$\widehat{K}_{ij} = \frac{\#\ Events\ originating\ from\ bin\ j\ that\ have\ been\ recorded\ by\ detector\ in\ bin\ i}{\#\ Events\ originating\ from\ bin\ j}$$

- The response matrix can be noisy, especially with a small MC sample size.



Two-step Approach

- Recall that

$$K_{ij} = \frac{\int_{y \in S_i} \int_{x \in T_j} k(y,x) f(x) dx dy}{\int_{x \in T_j} f(x) dx}$$

- Consider the estimator

$$\hat{K}_{ij} = \frac{\int_{y \in S_i} \int_{x \in T_j} \hat{k}(y,x) f(x) dx dy}{\int_{x \in T_j} f(x) dx}$$

1. Estimate the response kernel k on the unbinned space.
2. Plug back into the above equation.

- Potentially provide smoother estimate for K_{ij}.

Response kernel estimation

- $k(y, x) = p(smeared\ observation\ y \mid true\ event\ x)$.
- Given $(X_1, Y_1), \ldots, (X_n, Y_n) \sim p_{X,Y}$ from Monte Carlo generator, where X_i denotes the particle-level data and Y_i denotes the detector-level observation, estimating $k(y, x)$ is equivalent to conditional density estimation of $p_{Y|X}(y|x)$.
- Accurate estimate of the response kernel k should lead to accurate estimate of response matrix K.
- We will consider several nonparametric methods for conditional density estimation and make some comparisons.

Response kernel estimation

1. Kernel method

$$\hat{p}_{h_1,h_2}(y|x) = argmin_a \sum_{i=1}^{n} \left(K_{h_2}(y - Y_i) - a\right)^2 K_{h_1}(x - X_i)$$

$$= \sum_{i=1}^{n} w_i(x) K_{h_2}(y - Y_i)$$

where $w_i(x) = \frac{K_{h_1}(x - X_i)}{\sum_{j=1}^{n} K_{h_1}(x - X_j)}$ and K_h is some kernel function with bandwidth h > 0 (not the response kernel).

2. Local linear method

$$(\hat{a}, \hat{b}) = argmin_{a,b} \sum_{i=1}^{n} \left(K_{h_2}(y - Y_i) - a - b(X_i - x)\right)^2 K_{h_1}(x - X_i)$$

$$\hat{p}_{h_1,h_2}(y|x) = \hat{a}$$

- Two global bandwidth parameters h_1, h_2 control the amount of smoothing along X and Y, respectively.

Response kernel estimation

- Global bandwidth is not optimal in some cases, e.g. different amount of smearing applied to different regions for the response matrix.

3. Kernel method with local bandwidths

$$\hat{p}_{h_1(x),h_2(x)}(y|x) = \sum_{i:||x-X_i||<\delta(x)} w_i(x) K_{h_2(x)}(y - Y_i)$$

where $w_i(x) = \frac{K_{h_1(x)}(x-X_i)}{\sum_{j:||x-X_j||} K_{h_1(x)}(x-X_j)}$ and $\delta(x)$ is the window size at x.

- Local bandwidth parameters $h_1(x), h_2(x)$ control the amount of smoothing along X and Y conditioning on each x.

Response kernel estimation

4. Location-scale model

Suppose we assume the smeared observations are generated from the following model

$$Y = \mu(X) + \sigma(X)\epsilon$$

where ϵ follows some distribution with mean 0 and variance 1.

- Then $p(y|x)$ can be written as

$$p(y|x) = \frac{1}{\sigma(x)} p_\epsilon\left(\frac{y - \mu(x)}{\sigma(x)}\right)$$

and an estimator can be obtained by

$$\hat{p}(y|x) = \frac{1}{\hat{\sigma}(x)} \hat{p}_\epsilon\left(\frac{y - \hat{\mu}(x)}{\hat{\sigma}(x)}\right).$$

- $\hat{\mu}, \hat{\sigma}^2$ can be estimates by some regression method (e.g. splines) and $\hat{p}_\epsilon$ by density estimation (e.g. KDE).
- Directly model the variance function $\sigma^2(x)$ and hence avoid the problem of finding local bandwidths as in the case of local kernel method.

Simulation study

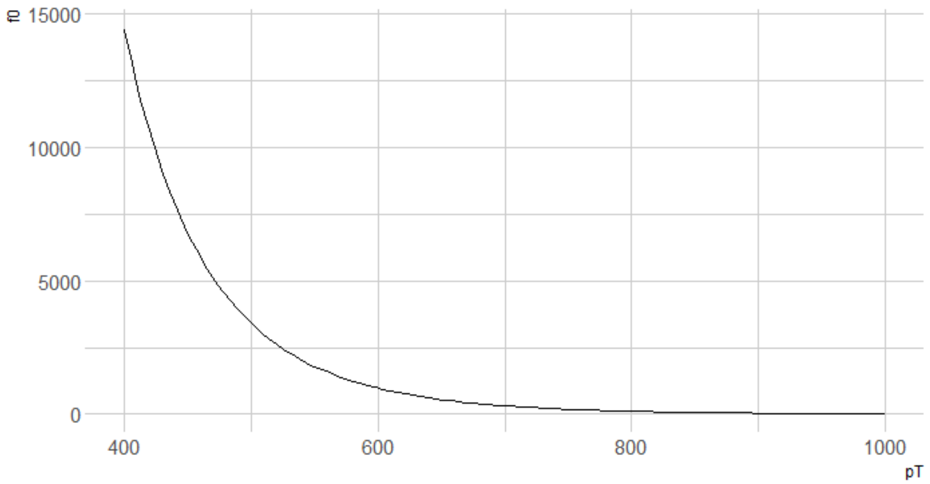
- We mimic unfolding the inclusive jet transverse momentum spectrum by simulating the data using the particle-level function

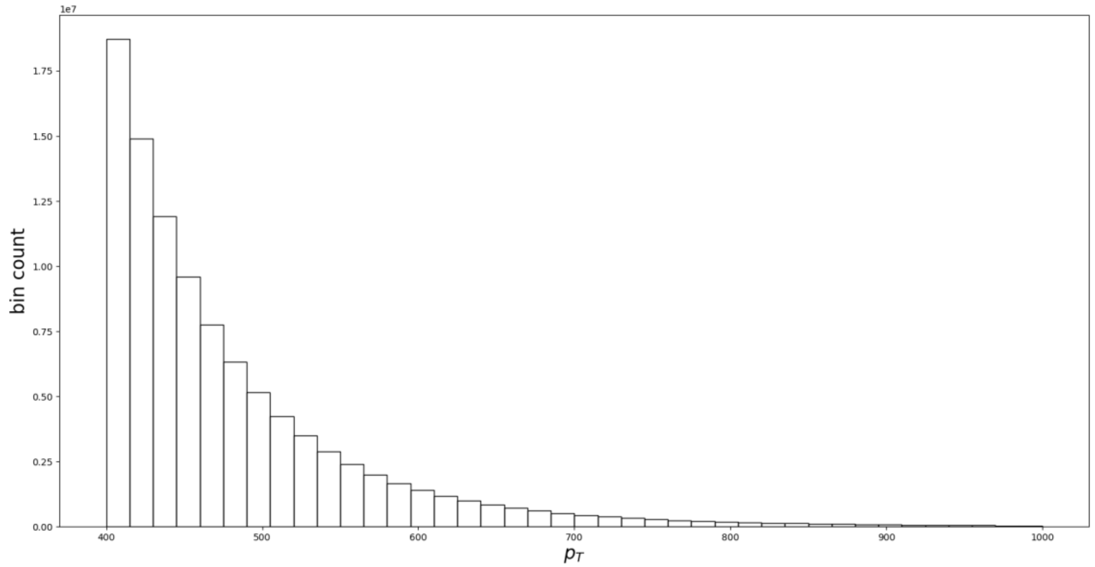
$$f(p_\perp) = LN_0\left(\frac{p_\perp}{GeV}\right)^{-\alpha}\left(1-\frac{2}{\sqrt{s}}p_\perp\right)^{\beta} e^{-\gamma/p_\perp}$$

- The parameters are given by

$L = 5.1fb^{-1}, N_0 = 10^{17}\frac{fb}{GeV}, \alpha = 5, \beta = 10, \gamma = 10\ GeV, \sqrt{s} = 7\ TeV.$

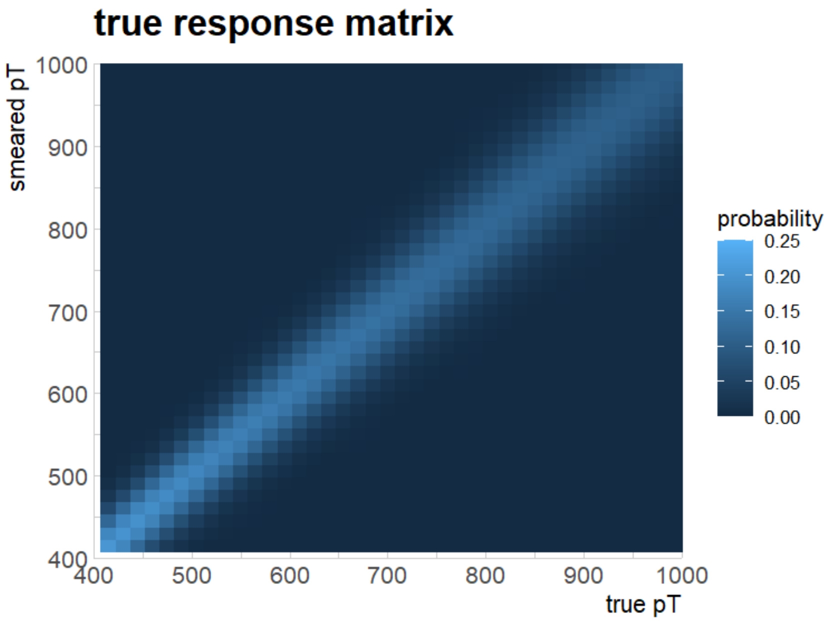
- The number of bins = 40.





Simulation study

- The response kernel is modeled as an additive Gaussian noise
$$k(p'_\perp, p_\perp) = N(p'_\perp - p_\perp | 0, \sigma(p_\perp)^2)$$
with heteroscedastic variance satisfying
$$\left(\frac{\sigma(p_\perp)}{p_\perp}\right)^2 = \left(\frac{C_1}{\sqrt{p_\perp}}\right)^2 + \left(\frac{C_2}{p_\perp}\right)^2 + C_3^2.$$
- The parameters are $C_1 = 1GeV^{1/2}, C_2 = 1GeV, C_3 = 0.05$.

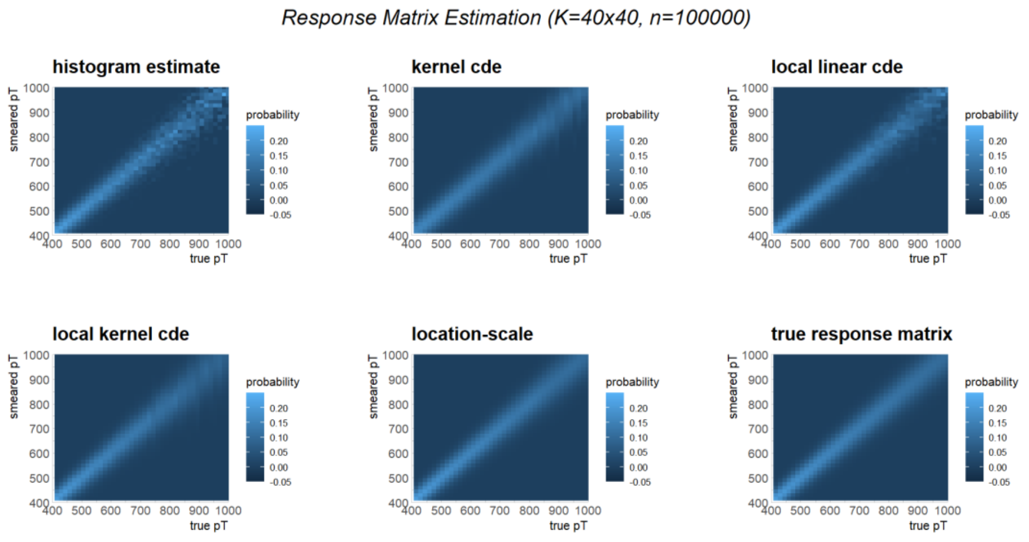


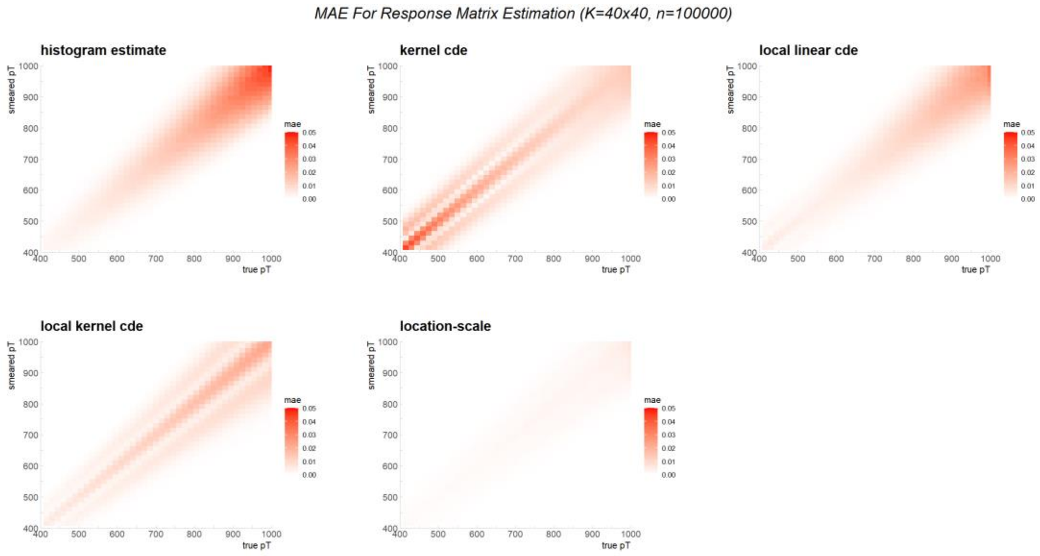
Comparison of the response matrix estimators

- The sample size (number of paired Monte Carlo events) for estimating the response matrix K is 100000.
- The performance of the estimators is compared using bin-wise mean absolute error (MAE)

$$\frac{1}{M}\sum_{l=1}^{M}\left|\widehat{K}_{ij}^{(l)} - K_{ij}\right| \quad for\ all\ i \in [m], j \in [n]$$

with $M = 1000$ Monte Carlo simulations.





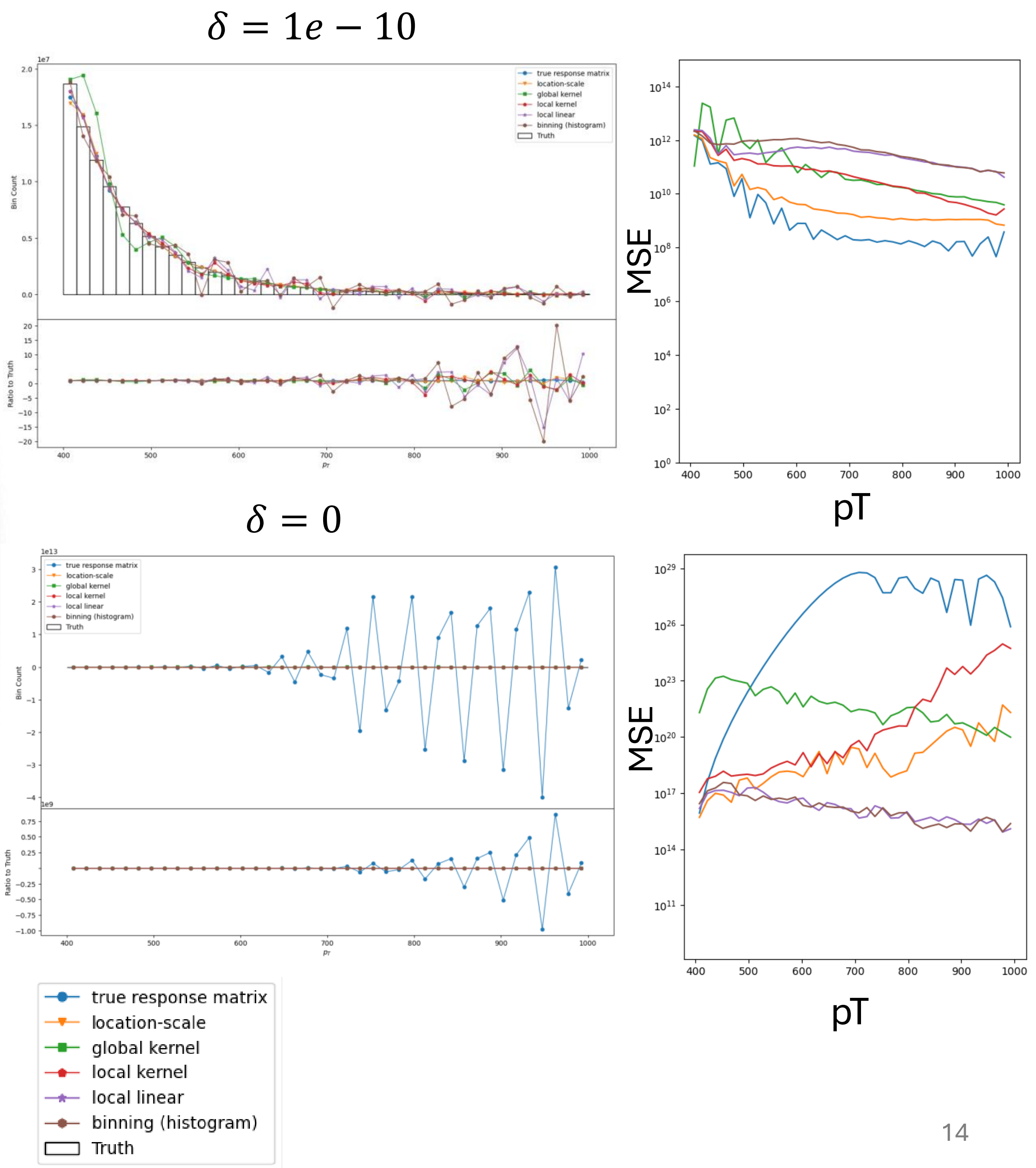
Effect of the estimated response matrix on the unfolded spectrum

- Does a better estimated response matrix lead to a better unfolded point estimator?
- Least-squares estimator with Tikhonov regularization.
- D'Agostini iteration (EM algorithm, Iterative Bayesian unfolding, Lucy-Richardson deconvolution).

Tikhonov regularization

- With some $\delta \geq 0$, the least squares solution with Tikhonov regularization is
$$\hat{\boldsymbol{\lambda}} = \left(\widehat{\boldsymbol{K}}^{\top}\widehat{\boldsymbol{K}} + \delta \boldsymbol{I}\right)^{-1}\widehat{\boldsymbol{K}}^{\top}\boldsymbol{y}.$$
- Better estimated response matrix generally leads to better unfolded solution.
- When there is no regularization ($\delta = 0$), the solution with the true response matrix (without noise) performs worse compared to estimated response matrices.
- The estimated response matrices implicitly perform regularization (an ill-conditioned matrix with some additive random noise becomes well-conditioned with high probability[1]).

[1] T. Tao, V. Vu, The condition number of a randomly perturbed matrix, in: Symposium on the Theory of Computing, 2007.



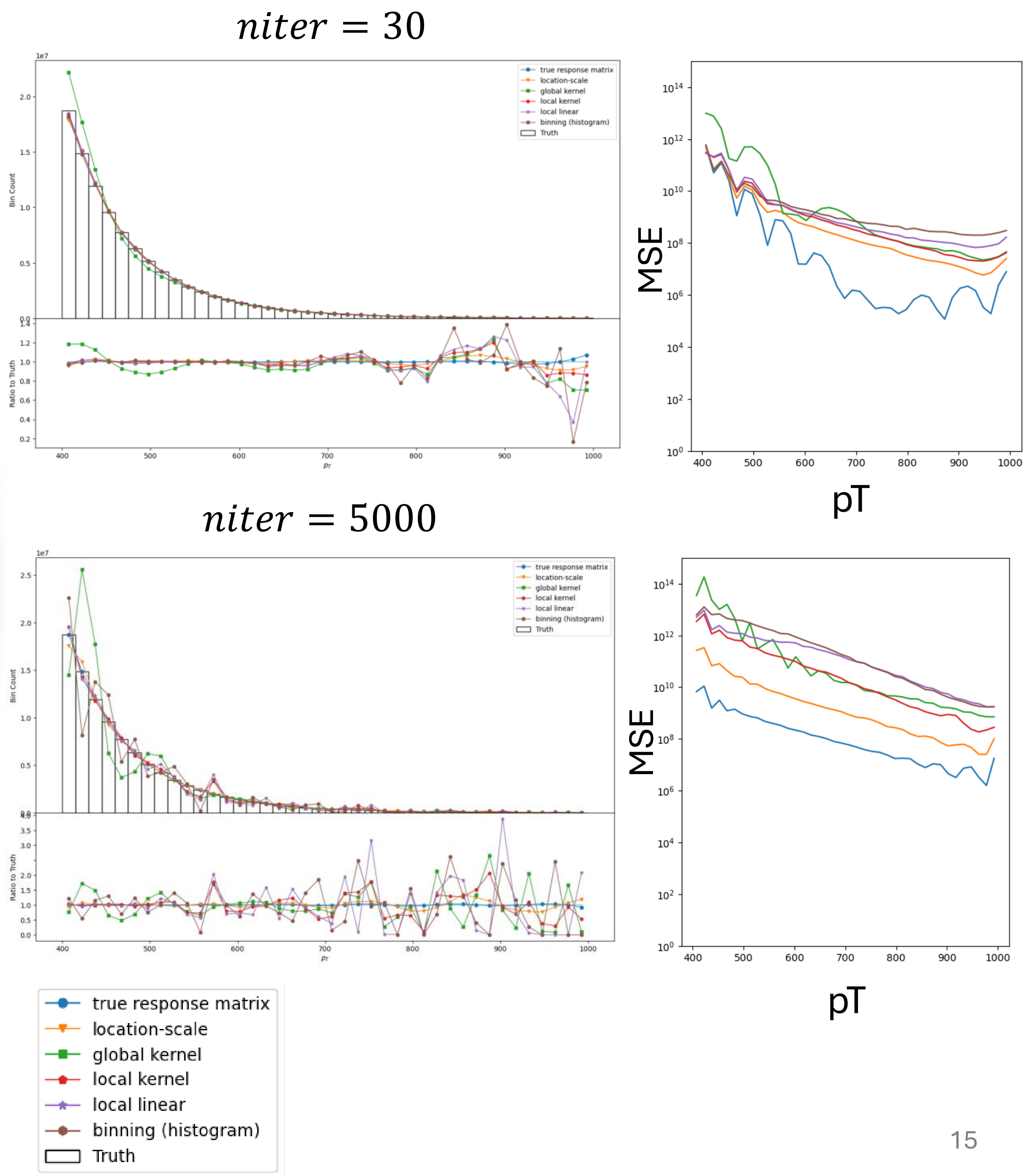
D’Agostini iteration

- After $r+1$ iterations, the solution is given by

$$\hat{\lambda}_j^{(r+1)} = \frac{\hat{\lambda}_j^{(r)}}{\sum_{i=1}^{m} \hat{K}_{ij}} \sum_{i=1}^{m} \frac{\hat{K}_{ij} y_i}{\sum_{l=1}^{n} \hat{K}_{il}\, \hat{\lambda}_l^{(r)}}$$

$$\hat{\boldsymbol{\lambda}}^{(r+1)} = \left[\hat{\lambda}_1^{(r+1)}, \ldots, \hat{\lambda}_n^{(r+1)}\right]$$

- Again, better estimated response matrix generally leads to better unfolded solution.
- Most estimated response matrices lead to similar MSE when the number of iterations is small.



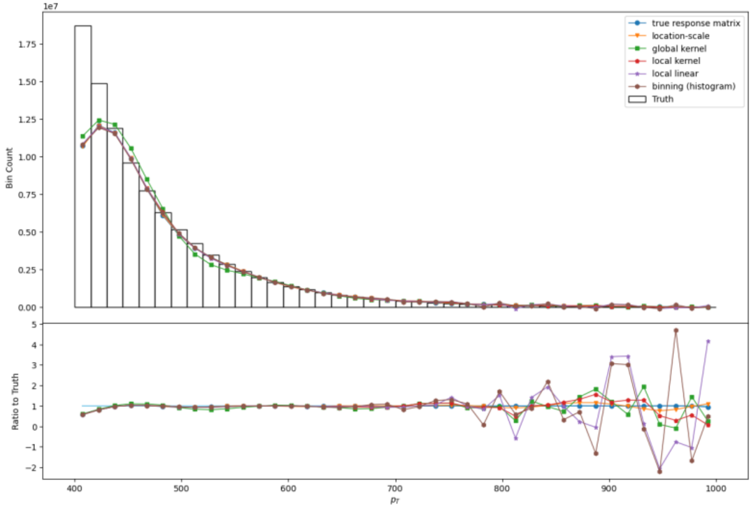
Summary

- Estimated response matrix from a Monte Carlo simulation has statistical uncertainty.
- Traditional binning (histogram) method can be noisy in regions that have small sample sizes.
- Two-step approach can remedy this issue by first estimating response kernel using conditional density estimation on the unbinned space, and then constructing a plug-in estimator of response matrix based on the estimated response kernel.
- The estimated response matrix is a more well-conditioned matrix compared to the true response matrix without any noise, which implicitly regularizes the solution.
- Uncertainty quantification for the unfolded solution in the presence of uncertainty in the response matrix is not immediately clear.

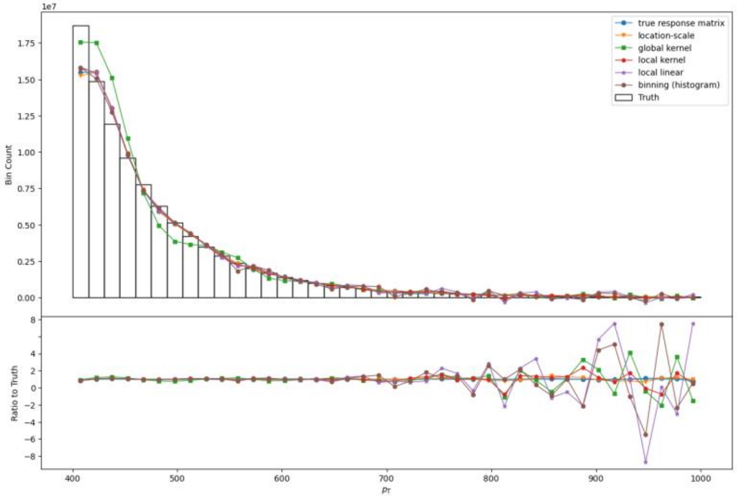
Backup

- Tikhonov regularization with different regularization strengths

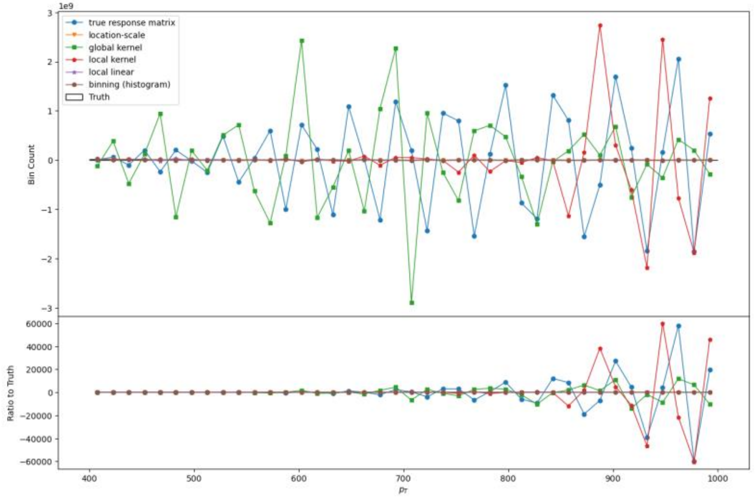
$\delta = 1e - 8$

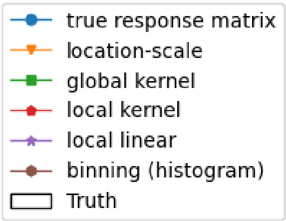


$\delta = 1e - 9$



$\delta = 1e - 20$

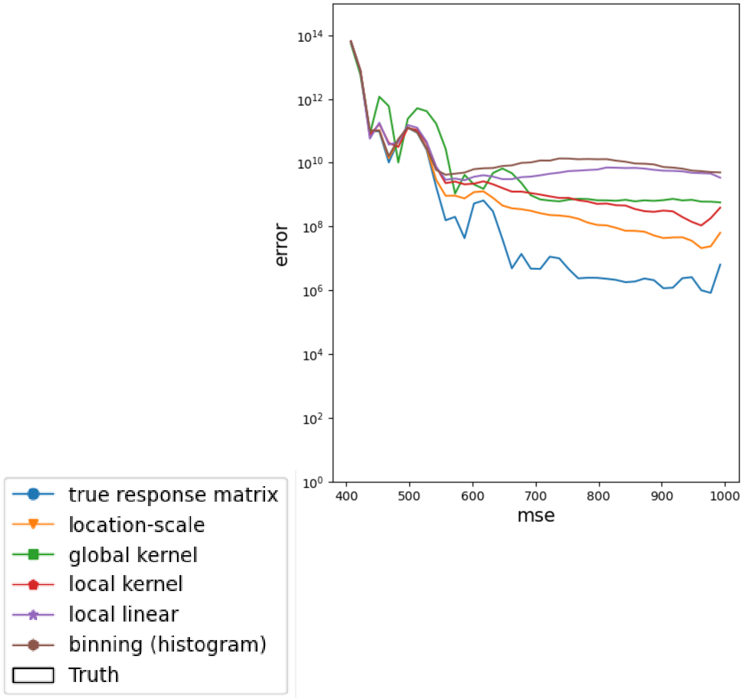




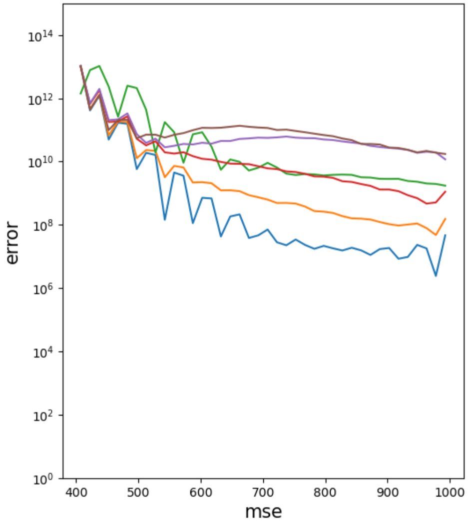
Backup

- MSE for Tikhonov regularization with different regularization strengths

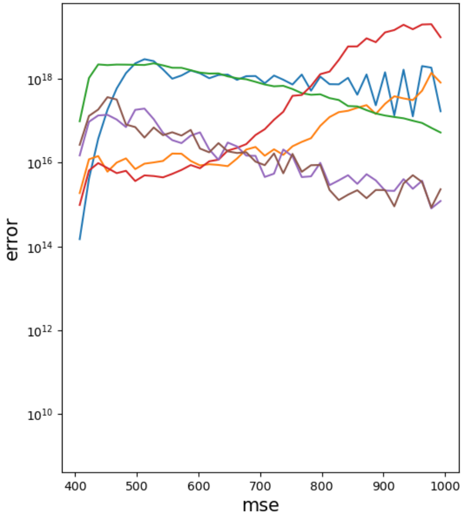
$\delta = 1e - 8$



$\delta = 1e - 9$



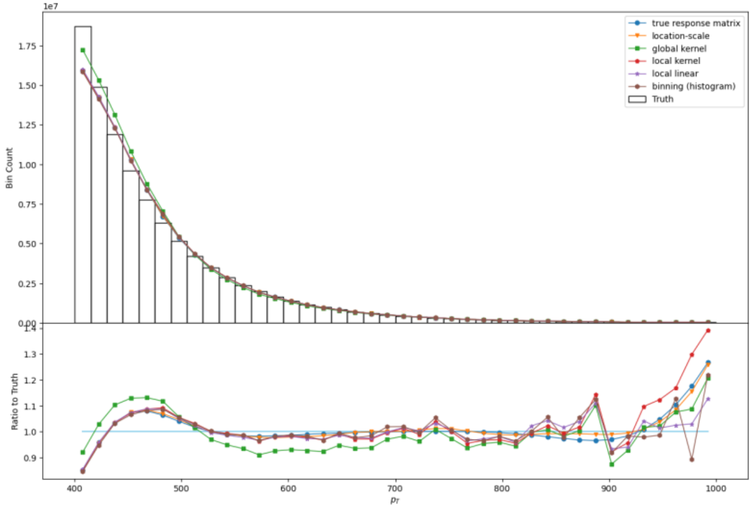
$\delta = 1e - 20$



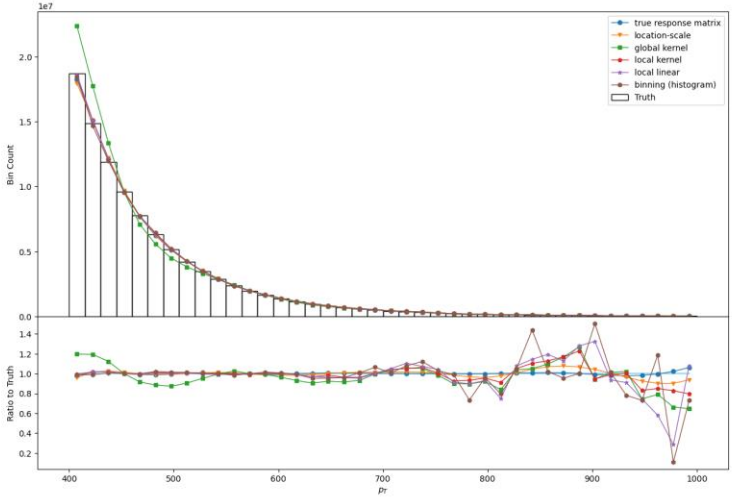
Backup

- D'Agostini solution with different number of iterations

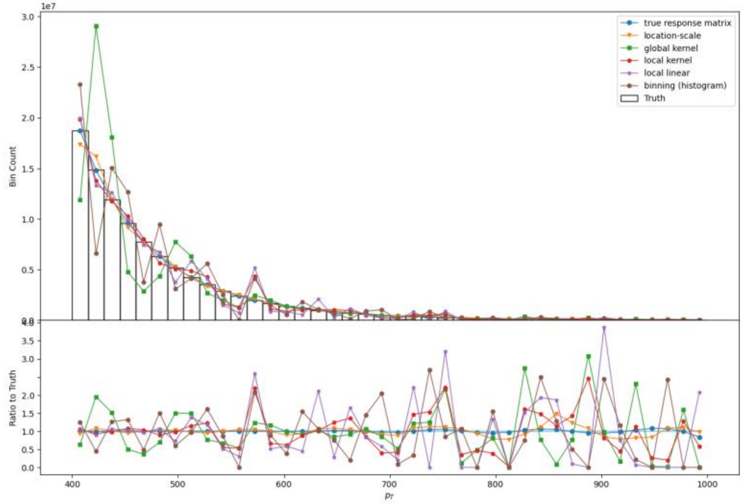
$niter = 3$

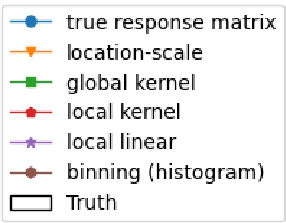


$niter = 40$



$niter = 10000$





Backup

- MSE for D’Agostini solution with different number of iterations

$niter = 3$ $niter = 40$ $niter = 10000$