Quantifying systematic uncertainty in unfolding forward models using optimal transport

Richard Zhu [1] Mikael Kuusela [1] Larry Wasserman [1]
Andrea Marini [2]

[1]Department of Statistics and Data Science, Carnegie Mellon University

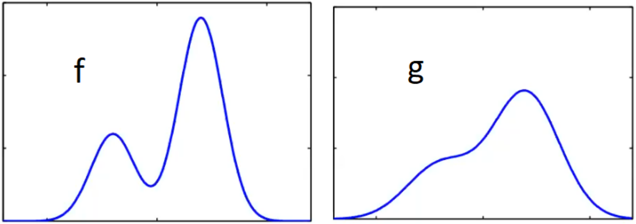
[2]CERN, the European Organization for Nuclear Research

27 April 2023

The unfolding problem: inferring the true particle spectrum from smeared observations

- In measurement analyses, one is interested in the distribution (spectrum) of some physical quantity, e.g., the energy, mass, momentum.

- Due to the finite resolution of the detectors, only a smeared version of the physical quantity is observed.



true distribution f smeared distribution g

Figure: unfolding [Kuusela (2016)]

Forward model for unfolding

Let f be the true distribution. The observed smeared distribution g is given by

$$g(s) = \int_T k(s,t) f(t) dt$$

where the response kernel k represents the response of the detector and is given by

$$k(s,t) = P(Y = s | X = t)$$

$X =$ true collision event and $Y =$ smeared observation.

Uncertainty in the forward model

- The response kernel $k(s, t)$ is usually not available in closed form and needs to be estimated using detector simulation.

- The imperfect knowledge of the detector alignment and calibration as well as the distribution of auxiliary variables can affect the response kernel in different ways.

- This leads to systematic uncertainty in the response kernel and hence the unfolded solution as well.

Using optimal transport to quantify uncertainty

- Given two kernels k_1, k_2, the 2-Wasserstein distance between k_1 and k_2 is defined as

$$W_2(k_1, k_2) = \left(\int_0^1 \left(F_1^{-1}(q) - F_2^{-1}(q) \right)^2 dq \right)^{1/2}$$

- F_1^{-1} is the quantile function of k_1 and F_2^{-1} is the quantile function of k_2 conditioned on a fixed t.

Using optimal transport to quantify uncertainty

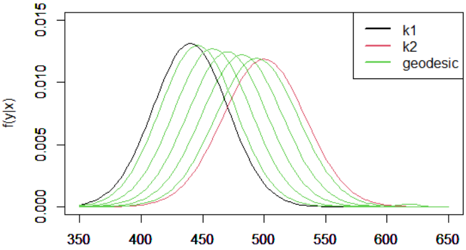
- The Wasserstein barycenter of k_1 and k_2 with weights $\mathbf{t} = (t_1, t_2)$ is given by

$$k_{\mathbf{t}} = \arg\min_k \{t_1 W_2(k_1, k) + t_2 W_2(k_2, k)\}$$

- More specifically, in 1d, the quantile function of $k_{\mathbf{t}}$ satisfies

$$F_{\mathbf{t}}^{-1} = t_1 F_1^{-1} + t_2 F_2^{-1}$$

- Varying the weight $\mathbf{t}$ defines the geodesic (path) morphing between k_1 and k_2: $\{k_{\mathbf{t}} : t_1, t_2 \geq 0, t_1 + t_2 = 1\}$.



Discretization

- Let $\{T_j\}_{j=1}^n$ be the partition of the true space T and $\{S_i\}_{i=1}^m$ be the partition of the smeared space.
- Particle-level histogram: $\mathbf{x} \sim \text{Poisson}(\boldsymbol{\lambda})$.
  Detector-level histogram $\mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu})$.
- True histogram mean: $\boldsymbol{\lambda} = [\int_{T_1} f(t)dt, ..., \int_{T_n} f(t)dt]$.
  Smeared histogram mean: $\boldsymbol{\mu} = [\int_{S_1} g(s)ds, ..., \int_{S_m} g(s)ds]$.
  f and g are the intensity functions of the Poisson processes.
- $\boldsymbol{\mu} = \mathbf{K}\boldsymbol{\lambda}$ where the elements of response matrix $\mathbf{K}$ are given by

$$\begin{aligned}\mathbf{K}_{ij} &= \frac{\int_{s\in S_i}\int_{t\in T_j} k(s,t)f(t)dtds}{\int_{t\in T_j} f(t)dt} \\ &= P\left(\text{smeared observation in bin i}|\text{true event in bin j}\right)\end{aligned}$$

Goal

Inference on the true histogram mean $\boldsymbol{\lambda}$.

Computing confidence interval for the true histogram mean while accounting for the systematic uncertainty in the response kernels

(1) Given two base kernels k_1, k_2, compute the geodesic $\{k_{\mathbf{t}} = \arg\min_k \{t_1 W(k_1, k) + t_2 W(k_2, k)\} : t_1, t_2 \geq 0, t_1 + t_2 = 1\}$.

(2) Compute the corresponding response matrices $\mathbf{K}_1, \mathbf{K}_2, \{\mathbf{K}_{\mathbf{t}}\}$.

(3) Unfold with One-at-a-time Strict-Bounds (OSB) (Stanley et al. (2022)) using the detector-level histogram $\mathbf{y}$ and response matrices $\mathbf{K}_1, \mathbf{K}_2, \{\mathbf{K}_{\mathbf{t}}\}$.

(4) Obtain a collection of confidence intervals $C_1, C_2, \{C_{\mathbf{t}}\}$ for $\boldsymbol{\lambda}$.

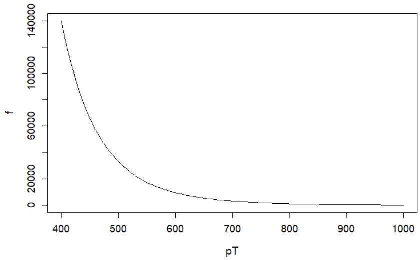
Simulation study - inclusive jet transverse momentum spectrum

- Simulate particle-level data using the intensity function

$$f_0\left(p_\perp\right) = LN_0\left(\frac{p_\perp}{\mathrm{GeV}}\right)^{-\alpha}\left(1-\frac{2}{\sqrt{s}}p_\perp\right)^{\beta}e^{-\gamma/p_\perp}, \quad 0 < p_\perp \leq \frac{\sqrt{s}}{2}$$

- Number of bins in detector level $= 40$
- Number of fine bins in particle level $= 40$
- Number of wide bins in particle level $= 10$

Simulation study - inclusive jet transverse momentum spectrum



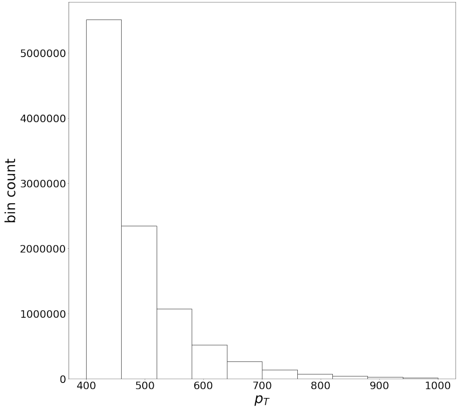


Figure: **LEFT**: intensity function; **RIGHT**: True histogram mean $\boldsymbol{\lambda}$

Simulation study - inclusive jet transverse momentum spectrum

- The detector smearing is modeled using crystal ball function

$$CB(t-s|\mu,\sigma,\alpha,\gamma) \propto \begin{cases} e^{\frac{-(t-s-\mu)^2}{2\sigma^2}} & \frac{t-s-\mu}{\sigma} > -\alpha \\ \left(\frac{\gamma}{\alpha}\right)^{\gamma} e^{-\frac{\alpha^2}{2}} \left(\frac{\gamma}{\alpha} - \alpha - \frac{t-s-\mu}{\sigma}\right)^{-\gamma} & \frac{t-s-\mu}{\sigma} \leq -\alpha \end{cases}$$

- Two base kernels

$$k_1 : \mu = 0, \sigma = 10, \alpha = 1, \gamma = 2$$

$$k_2 : \mu = 7, \sigma = 12, \alpha = 1, \gamma = 2$$

Simulation study - inclusive jet transverse momentum spectrum

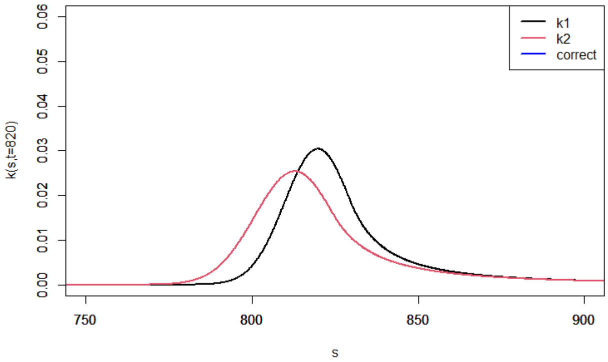


Figure: k1 and k2 are the base kernels that we might obtain from detector simulation

Simulation study - inclusive jet transverse momentum spectrum

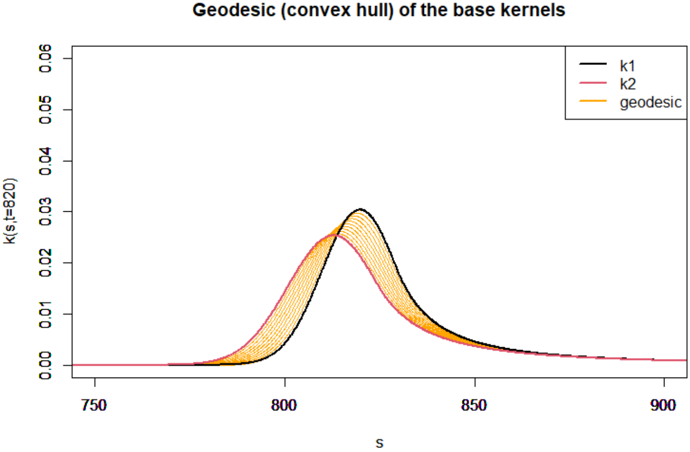


Figure: Wasserstein geodesic of k1 and k2

Simulation study - inclusive jet transverse momentum spectrum

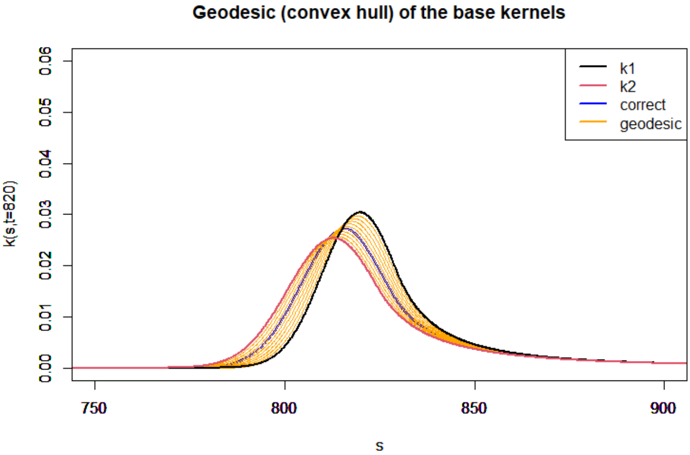


Figure: correct kernel represents the actual unknown detector response that generates the smeared observation

Unfold with the geodesic of the kernels

- We unfold with the geodesic of the kernels using the OSB intervals on one of the bins.

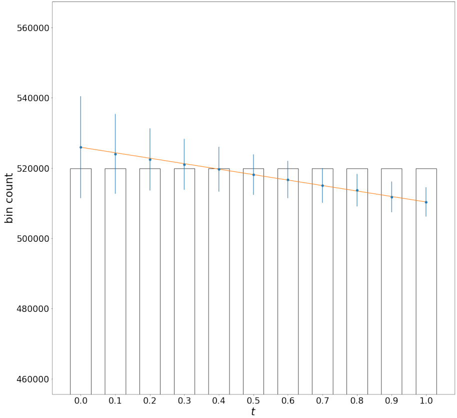


Figure: OSB confidence intervals for $\boldsymbol{\lambda}$ for bin 4; x-axis represents the weight that determines the kernel on the geodesic.

Unfold with the geodesic of the kernels

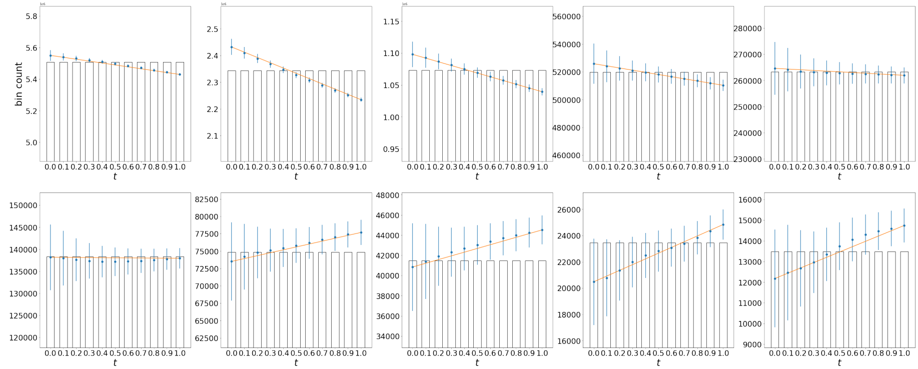


Figure: OSB confidence intervals for $\boldsymbol{\lambda}$ for 10 bins; each plot corresponds to 1 bin; x-axis represents the weight on the geodesic.

Unfold with the geodesic of the kernels

- We define *confidence slabs* to be the collection of 2-dimensional confidence sets for the true histogram mean $\boldsymbol{\lambda}$ of 2 bins unfolded by the geodesic of kernels defined by k_1 and k_2.
- Confidence slabs cover the true histogram mean.

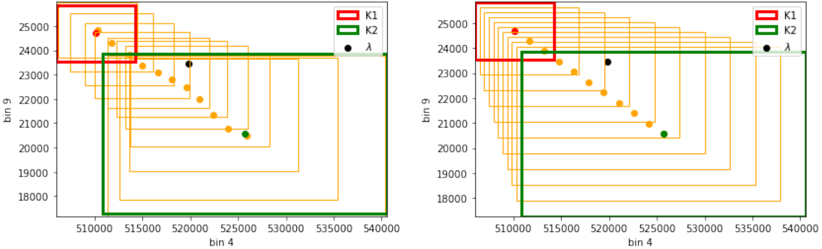


Figure: **LEFT**: Confidence slabs unfolded by the geodesic of K1 and K2; **RIGHT**: Interpolation of unfolded boxes by K1 and K2

Confidence slabs — more bins

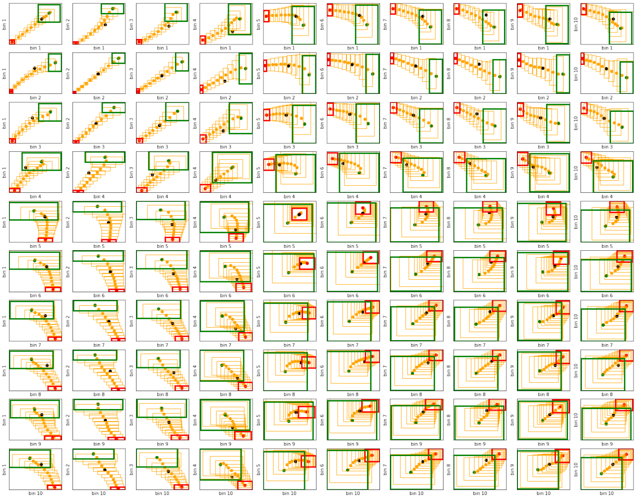


Figure: Confidence slabs for all bins. Presence of nonlinear patterns.

Confidence slabs have proper coverage

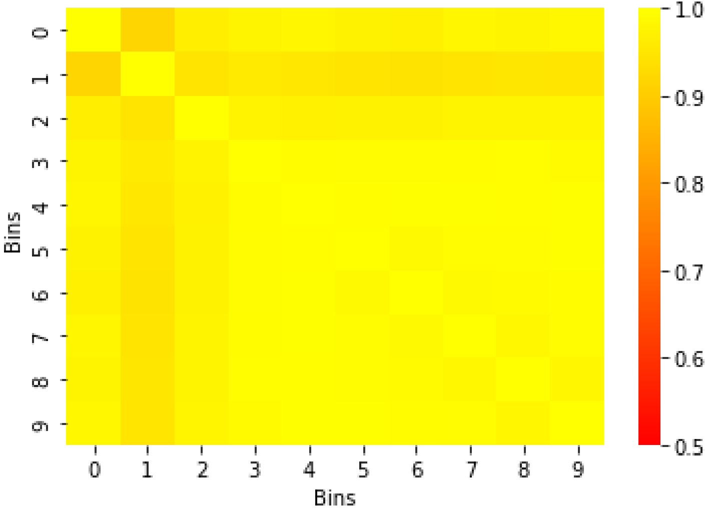


Figure: Coverage for confidence slabs

Extrapolation

- We can allow the weight $t_1, t_2 < 0$ to define extrapolation of the base kernels:

$$\{k_{\mathbf{t}} = \arg\min_k \{t_1 W(k_1, k) + t_2 W(k_2, k)\} : t_1 + t_2 = 1\}$$

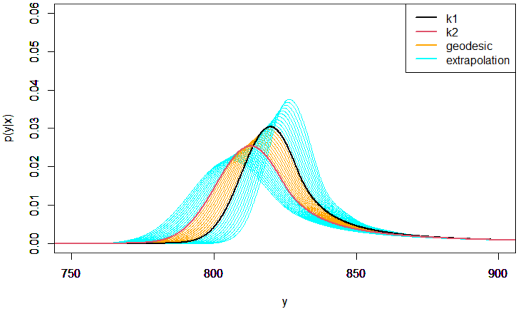


Figure: Extrapolation of base kernels

More base kernels

- The Wasserstein barycenter of $k_1, k_2, ..., k_m$ with weights $\mathbf{t} = (t_1, t_2, ...t_m)$ is given by

$$k_{\mathbf{t}} = \arg\min_k \{\sum_{i=1}^m t_i W_2(k_i, k)\}$$

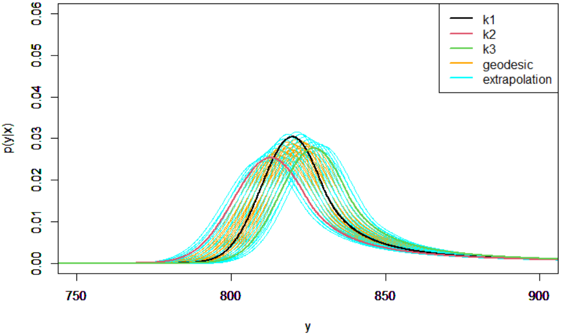
with quantile function

$$F_{\mathbf{t}}^{-1} = \sum_{i=1}^m t_i F_i^{-1}$$

- Varying the weight $\mathbf{t}$ defines the Wasserstein hull of kernels defined by $k_1, k_2, ..., k_m$: $\{k_{\mathbf{t}} : \sum_{i=1}^m t_i = 1\}$.

More base kernels

- Varying the weight $\mathbf{t}$ defines the Wasserstein hull of kernels defined by $k_1, k_2, ..., k_m$: $\{k_{\mathbf{t}} : \sum_{i=1}^{m} t_i = 1\}$.



Summary and Open Problems

- The unfolding problem: Systematic uncertainty in the forward model.
- Method: Use optimal transport to quantify the uncertainty in the response kernel.
- Results: Confidence slabs with proper coverage when the correct kernel is on (or close to) the geodesic of the base kernels.
- Open problems: For a given kernel $k_{\mathbf{t}}$ on the geodesic, we can view the weight $\mathbf{t}$ as a nuisance parameter. How can we summarize the collection of confidence intervals (dependent on $\mathbf{t}$) into a single confidence interval? Can we do profile likelihood? Can we learn $\mathbf{t}$ from the data? How well does it work on real HEP analysis?

Thank you!

Travel support from the NSF AI Planning Institute for Data-Driven Discovery in Physics is gratefully acknowledged.

APPENDIX: Simulation study - inclusive jet transverse momentum spectrum

- The detector smearing is modeled using crystal ball function

$$CB(t-s|\mu,\sigma,\alpha,\gamma) \propto \begin{cases} e^{\frac{(t-s-\mu)^2}{2\sigma^2}} & \frac{t-s-\mu}{\sigma} > -\alpha \\ \left(\frac{\gamma}{\alpha}\right)^{\gamma} e^{-\frac{\alpha^2}{2}} \left(\frac{\gamma}{\alpha} - \alpha - \frac{t-s-\mu}{\sigma}\right)^{-\gamma} & \frac{t-s-\mu}{\sigma} \leq -\alpha \end{cases}$$

- One correct kernel and two alternative kernels

$$k_{correct} : \mu = 3, \sigma = 11, \alpha = 1, \gamma = 2$$

$$k_1 : \mu = 0, \sigma = 10, \alpha = 1, \gamma = 2$$

$$k_2 : \mu = 10, \sigma = 12, \alpha = 1, \gamma = 2$$

Simulation study - inclusive jet transverse momentum spectrum

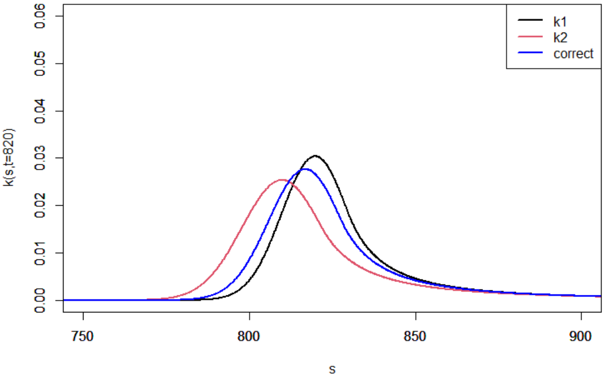


Figure: k1 and k2 are the base kernels that we might obtain from detector simulation; correct kernel represents the actual unknown detector response that generates the smeared observation.

Simulation study - inclusive jet transverse momentum spectrum

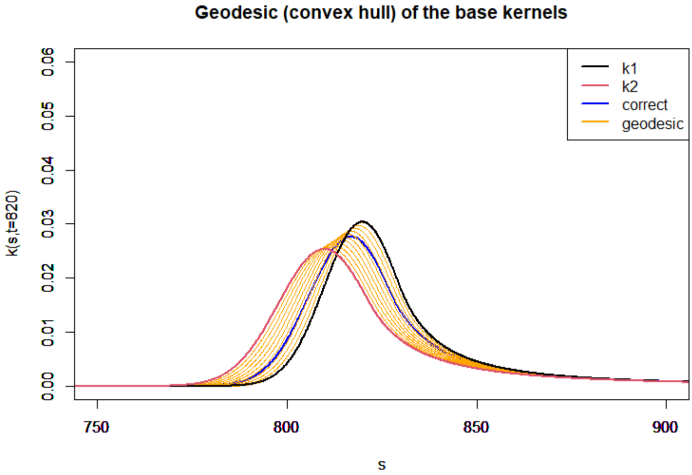


Figure: Wasserstein geodesic of k1 and k2

Unfold with the geodesic of the kernels

- We use the midpoints of the OSB intervals as the point estimates for $\boldsymbol{\lambda}$.

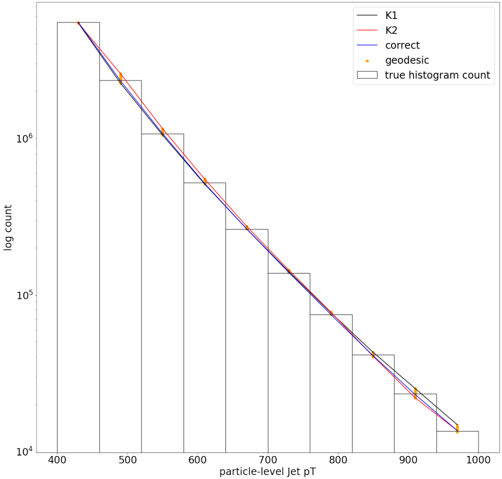
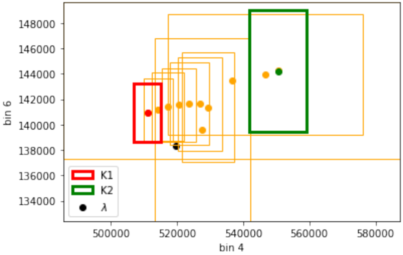


Figure: OSB midpoint solutions for geodesic of two kernels

Unfold with the geodesic of the kernels

- We define *confidence slabs* to be the collection of 2-dimensional confidence sets for the true histogram mean $\boldsymbol{\lambda}$ of 2 bins unfolded by the geodesic of kernels defined by k_1 and k_2.
- Confidence slabs cover the true histogram mean.
- The interpolation between the two corner confidence boxes (unfolded by k_1 and k_2) fails to cover the true mean.



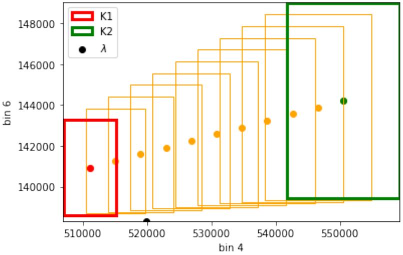


Figure: **LEFT**: Confidence slabs unfolded by the geodesic of K1 and K2; **RIGHT**: Interpolation of unfolded boxes by K1 and K2

Confidence slabs — more bins

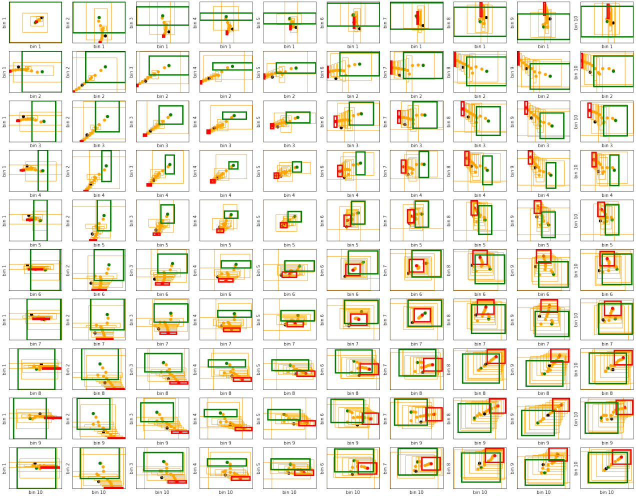


Figure: Confidence slabs for all bins. Presence of nonlinear patterns.

Confidence slabs have proper coverage

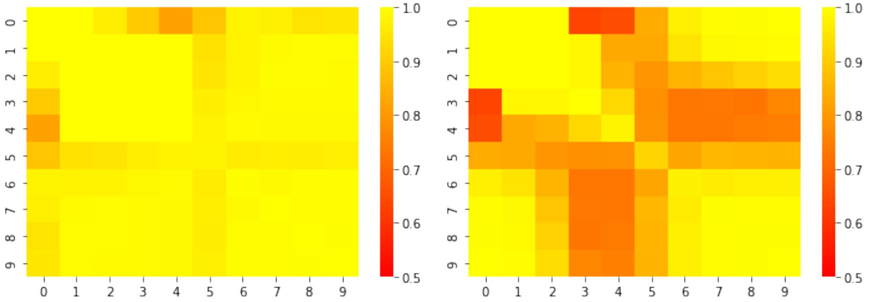


Figure: **LEFT**: Coverage for confidence slabs; **RIGHT**: Coverage for interpolation

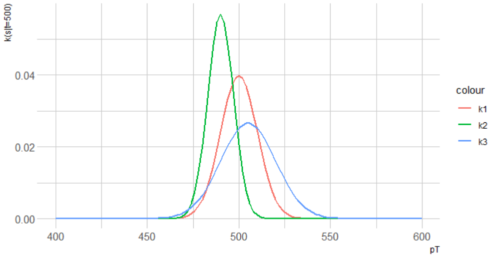
Simulation study - inclusive jet transverse momentum spectrum

- The detector smearing is modeled using Gaussian kernel

$$k_1(s,t) = N(s|\mu = t, \sigma = 10) \quad \text{(correct)}$$

- Two alternative kernels

$$k_2(s,t) = N(s|\mu = 0.98t, \sigma = 7), k_3(s,t) = N(s|\mu = 1.01t, \sigma = 15)$$



Unfold with the geodesic of the kernels

- We use the midpoints of the OSB intervals as the point estimates for $\boldsymbol{\lambda}$.

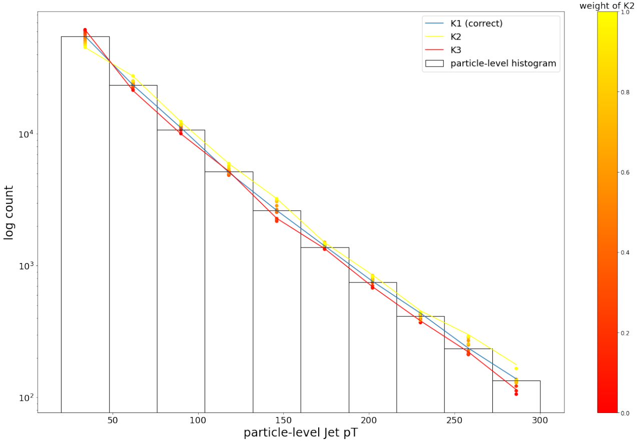


Figure: OSB midpoint solutions for geodesic of two kernels

Unfold with the geodesic of the kernels

- We define *confidence slabs* to be the collection of 2-dimensional confidence sets for the true histogram mean $\boldsymbol{\lambda}$ of 2 bins unfolded by the geodesic of kernels defined by k_2 and k_3.
- Confidence slabs cover the true histogram mean.
- The range of the confidence slabs is much smaller compared to the span of the confidence sets unfolded by the two corner kernels k_2, k_3 ("two-point" confidence sets).

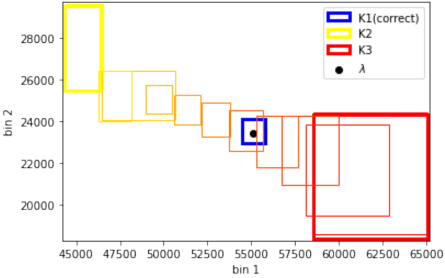


Figure: Confidence slab for bin 1 and bin 2

Confidence slabs — more bins

Figure: Confidence slabs for the first 5 bins

Confidence slabs have proper coverage

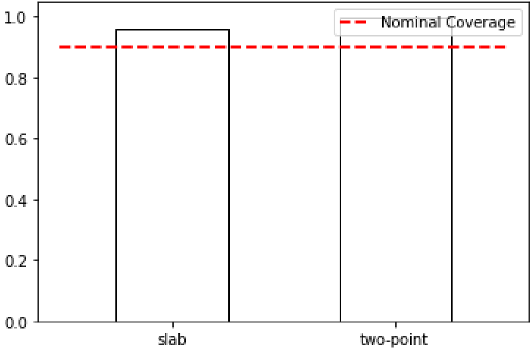


Figure: Coverage for confidence slabs and two-point confidence sets

Confidence slabs have proper coverage

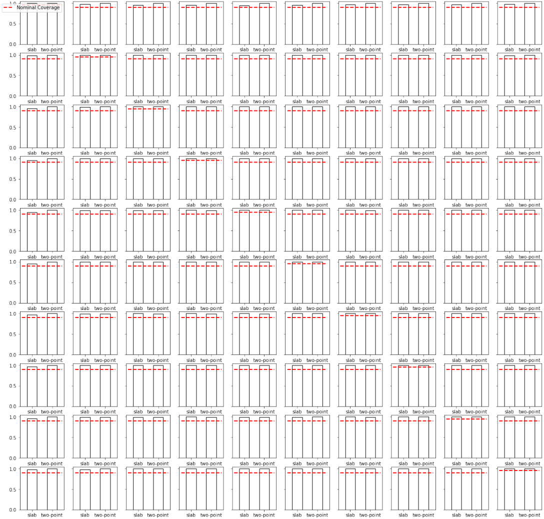


Figure: Coverage for confidence slabs and two-point confidence sets

Applications to simulated LHC data

- Unfold the jet transverse momentum spectrum in Drell-Yan events.
- Generate Monte Carlo events $\{X_i, Y_i\}_{i=1}^{n} \in \mathbb{R}^2 (n = 68180)$ corresponding to particle and detector level jet $p_\perp$ respectively.
- To produce alternative kernels, we simulate the effect of a jet energy uncertainty by location shifting and smearing of Y_i.

$$
\begin{aligned}
Y_i^{(1)} &= 1.02 Y_i + N(\mu = 0, sd = 10) \qquad\qquad \text{(correct)} \\
Y_i^{(2)} &= 1.1 Y_i + N(\mu = 0, sd = 20) \\
Y_i^{(3)} &= 0.9 Y_i + N(\mu = 0, sd = 5)
\end{aligned}
$$

- Obtain kernel estimates $\widehat{k}_1, \widehat{k}_2, \widehat{k}_3$ corresponding to $\{X_i, Y_i^{(1)}\}, \{X_i, Y_i^{(2)}\}, \{X_i, Y_i^{(3)}\}$.

Kernel estimation

- Kernel is the conditional density of smeared Y given true X: $k(y,x) = p(y|x)$.
- We assume

$$Y = m(x) + \sigma(x)\epsilon, \qquad \sigma(x) > 0, \epsilon \sim D(\mu = 0)$$

- Regress Y on X to obtain estimates $\widehat{m}(x)$ and residuals $\widehat{r}_i = y_i - \widehat{m}(x_i)$.
- Regress $\widehat{r}_i^2$ on x_i to obtain estimates $\widehat{\sigma}^2(x)$.
- Estimate the density of ϵ using $\frac{\widehat{r}_i}{\widehat{\sigma}(x_i)}$ and obtain $\widehat{p}_\epsilon$.
- Estimate the conditional density of Y given X by

$$\widehat{p}(y|x) = \frac{1}{\widehat{\sigma}(x)}\widehat{p}_\epsilon\left(\frac{y - \widehat{m}(x)}{\widehat{\sigma}(x)}\right)$$

Unfolding applied to simulated LHC data

- Perform the same unfolding procedure as in the simulation study, except we have estimated response kernels $\widehat{k}$, particle-level intensity function $\widehat{f}$.

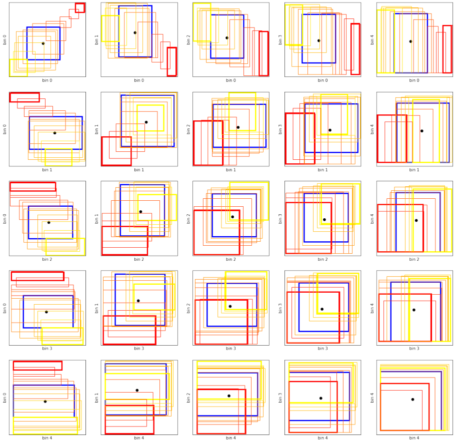


Figure: Confidence slabs for the first 5 bins

Unfolding applied to simulated LHC data

- In some cases, the confidence slabs (and the correct solution) can go outside the two-point confidence sets.

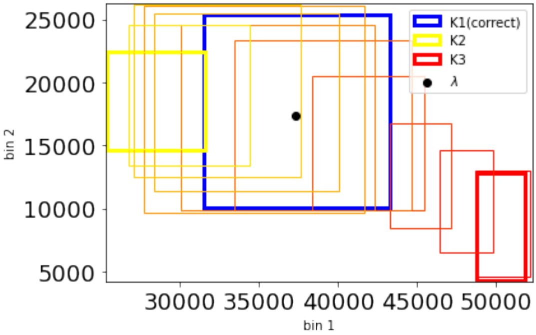
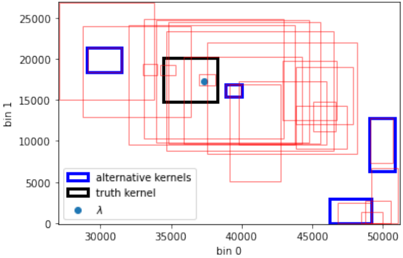


Figure: Confidence slabs for bin 1 and bin2

Unfolding with more kernels



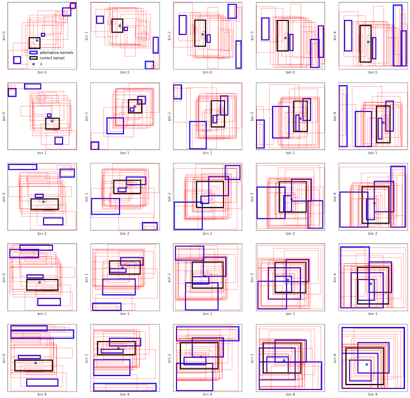


Figure: **LEFT**: Confidence slabs for bin 0 and bin 1; **RIGHT**: Confidence slabs for 5 bins