Unfolding: A Statistician's Perspective

Mikael Kuusela

Department of Statistics and Data Science
Carnegie Mellon University

NuXTract 2023
CERN, Geneva, Switzerland

October 3, 2023

The unfolding problem

- Any differential cross section measurement is affected by the finite resolution of the particle detectors
  - This causes the observed spectrum of events to be “smeared” or “blurred” with respect to the true one
- The *unfolding problem* is to estimate the true spectrum using the smeared observations
- Ill-posed inverse problem with major methodological challenges

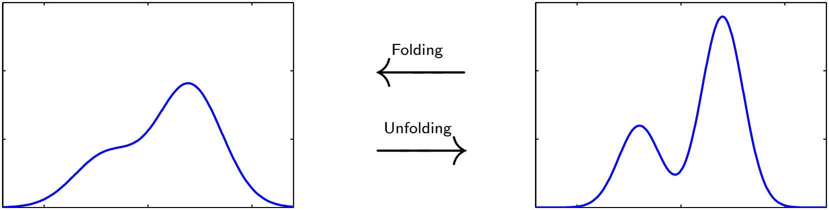


Figure: Smeared spectrum

Figure: True spectrum

Problem formulation

- Let f be the true, particle-level spectrum and g the smeared, detector-level spectrum
  - Denote the true space by T and the smeared space by S (both taken to be intervals on the real line for simplicity)
  - Mathematically f and g are the intensity functions of the underlying Poisson point process
- The two spectra are related by

$$g(s) = \int_T k(s,t) f(t) \, \mathrm{d}t,$$

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

$$k(s,t) = p(Y = s | X = t, X \text{ observed}) P(X \text{ observed} | X = t),$$

where X is a true event and Y the corresponding smeared event

Task: Infer the true spectrum f given smeared observations from g

Discretization

- Problem usually discretized using histograms (splines are also sometimes used)
- Let $\{T_i\}_{i=1}^p$ and $\{S_i\}_{i=1}^n$ be binnings of the true space T and the smeared space S
- Smeared histogram $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ with mean

$$\boldsymbol{\mu} = \left[\int_{S_1} g(s)\,\mathrm{d}s, \ldots, \int_{S_n} g(s)\,\mathrm{d}s \right]^{\mathrm{T}}$$

- Quantity of interest:

$$\boldsymbol{\lambda} = \left[\int_{T_1} f(t)\,\mathrm{d}t, \ldots, \int_{T_p} f(t)\,\mathrm{d}t \right]^{\mathrm{T}}$$

- The mean histograms are related by $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$, where the elements of the *response matrix* $\boldsymbol{K}$ are given by

$$K_{i,j} = \frac{\int_{S_i} \int_{T_j} k(s,t) f(t)\,\mathrm{d}t\,\mathrm{d}s}{\int_{T_j} f(t)\,\mathrm{d}t} = P(\text{smeared event in bin } i \,|\, \text{true event in bin } j)$$

- The discretized statistical model becomes

$$\boldsymbol{y} \sim \mathrm{Poisson}(\boldsymbol{K}\boldsymbol{\lambda})$$

and we wish to make inferences about $\boldsymbol{\lambda}$ under this model

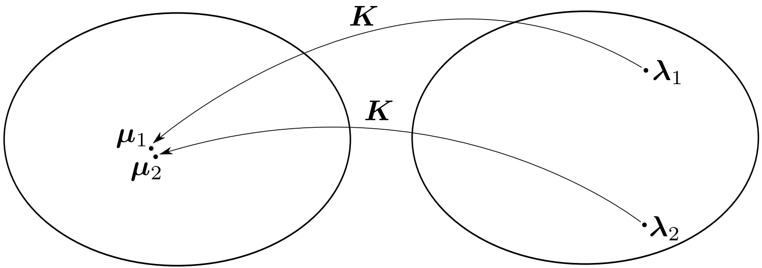
Why is unfolding difficult?

Two key challenges:

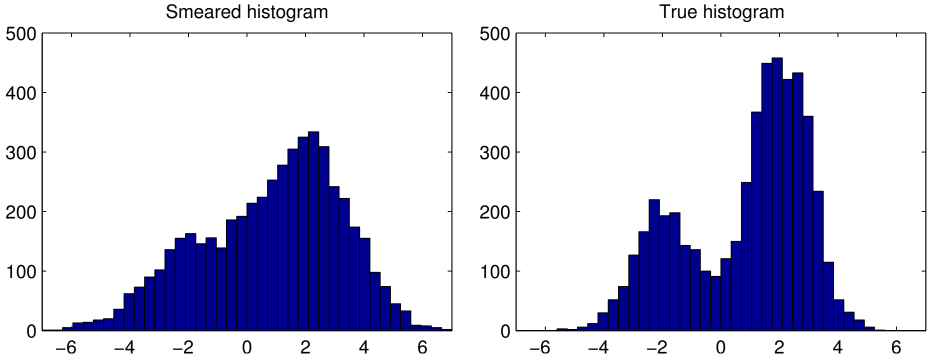
- Problem 1: $\boldsymbol{K}$ is an ill-conditioned matrix $\Rightarrow \hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$ tends to have unphysical high-frequency oscillations $\Rightarrow$ Regularization
- Problem 2: $\boldsymbol{K}$ depends on the shape of the spectrum inside the true bins $\Rightarrow$ $\boldsymbol{K}$ estimated using a MC ansatz $\Rightarrow$ Systematic uncertainty

Unfolding is an ill-posed inverse problem

- When the linear system $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ is ill-conditioned, true histograms $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ that are very different can map into smeared histograms $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ that are very similar
- As a result, distinguishing between $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ based on noisy data in the $\boldsymbol{\mu}$-space is very difficult

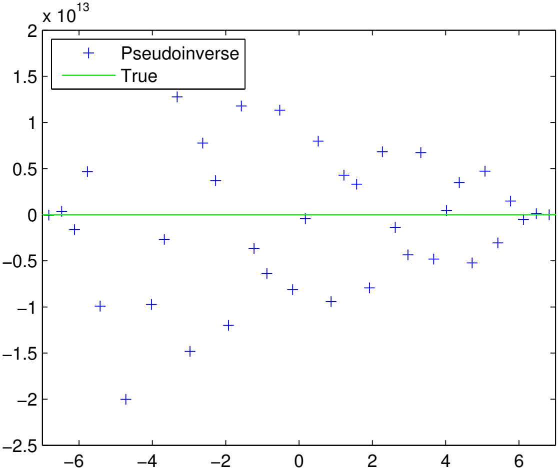


Demonstration of ill-posedness

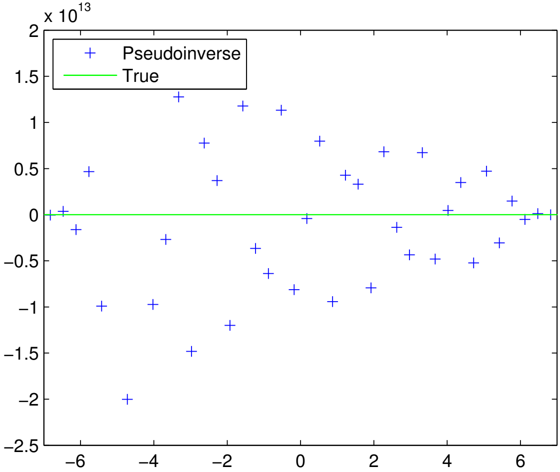


$$\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}, \quad \boldsymbol{y} \sim \mathrm{Poisson}(\boldsymbol{\mu}) \quad \overset{??}{\Longrightarrow} \quad \hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$$

Demonstration of ill-posedness



Demonstration of ill-posedness



$$\text{MSE}(\hat{\theta}) = \mathsf{E}((\hat{\theta} - \theta)^2) = [\text{bias}(\hat{\theta})]^2 + \mathsf{var}(\hat{\theta})$$

Regularization: bias $\uparrow$, variance $\downarrow$ $\Rightarrow$ $\text{MSE} \downarrow$

Regularization

Two main approaches to regularization:

1. Explicit penalty term
   - Tikhonov regularization / SVD unfolding / `TUnfold` (Höcker and Kartvelishvili, 1996; Schmitt, 2012)
2. Early stopping of an iterative algorithm
   - EM iteration with early stopping / D'Agostini iteration (D'Agostini, 1995; Richardson, 1972; Lucy, 1974; Shepp and Vardi, 1982; Lange and Carson, 1984; Vardi et al., 1985)

Tikhonov regularization

- Tikhonov regularization estimates $\boldsymbol{\lambda}$ by solving:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\lambda})^{\mathrm{T}} \hat{\boldsymbol{C}}^{-1} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\lambda}) + \delta P(\boldsymbol{\lambda})$$

- The first term as a Gaussian approximation to the Poisson log-likelihood
- The second term penalizes physically implausible solutions
- Common penalty terms:
  - Norm: $P(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|^2$
  - Curvature: $P(\boldsymbol{\lambda}) = \|\mathsf{L}\boldsymbol{\lambda}\|^2$, where L is a discretized 2nd derivative operator
  - SVD unfolding (Höcker and Kartvelishvili, 1996):

$$P(\boldsymbol{\lambda}) = \left\| \mathsf{L} \begin{bmatrix} \lambda_1 / \lambda_1^{\mathrm{MC}} \\ \lambda_2 / \lambda_2^{\mathrm{MC}} \\ \vdots \\ \lambda_p / \lambda_p^{\mathrm{MC}} \end{bmatrix} \right\|^2,$$

    where $\boldsymbol{\lambda}^{\mathrm{MC}}$ is a MC prediction for $\boldsymbol{\lambda}$
  - `TUnfold`[1] (Schmitt, 2012): $P(\boldsymbol{\lambda}) = \|\mathsf{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\mathrm{MC}})\|^2$

[1] `TUnfold` implements also more general penalty terms

D'Agostini iteration

- Starting from some initial guess $\boldsymbol{\lambda}^{(0)} > 0$, iterate

$$\lambda_j^{(k+1)} = \frac{\lambda_j^{(k)}}{\sum_{i=1}^{n} K_{i,j}} \sum_{i=1}^{n} \frac{K_{i,j} y_i}{\sum_{l=1}^{p} K_{i,l} \lambda_l^{(k)}}$$

- Regularization by stopping the iteration before convergence:
  - $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^{(K)}$ for some small number of iterations K
  - Will bias the solution towards $\boldsymbol{\lambda}^{(0)}$
  - Regularization strength controlled by the choice of K
- In `RooUnfold` (Adye, 2011), $\boldsymbol{\lambda}^{(0)} = \boldsymbol{\lambda}^{\mathrm{MC}}$
- `PyUnfold` (Bourbeau and Hampel-Arias, 2018) implements free choice of $\boldsymbol{\lambda}^{(0)}$

D'Agostini iteration

$$\lambda_j^{(k+1)} = \frac{\lambda_j^{(k)}}{\sum_{i=1}^{n} K_{i,j}} \sum_{i=1}^{n} \frac{K_{i,j} y_i}{\sum_{l=1}^{p} K_{i,l} \lambda_l^{(k)}}$$

- This iteration has been discovered in various fields, including optics (Richardson, 1972), astronomy (Lucy, 1974) and tomography (Shepp and Vardi, 1982; Lange and Carson, 1984; Vardi et al., 1985)
- In particle physics, it was popularized by D'Agostini (1995) who called it "Bayesian" unfolding
- But: This is in fact an expectation-maximization (EM) iteration (Dempster et al., 1977) for finding the *maximum likelihood estimator* of $\boldsymbol{\lambda}$ in the Poisson regression problem $\boldsymbol{y} \sim \mathrm{Poisson}(\boldsymbol{K}\boldsymbol{\lambda})$
- As $k \to \infty$, $\boldsymbol{\lambda}^{(k)} \to \hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}$ (Vardi et al., 1985)
- *This is a fully frequentist technique for finding the (regularized) MLE*
    - The name "Bayesian" is an unfortunate misnomer

D'Agostini demo, $k = 0$

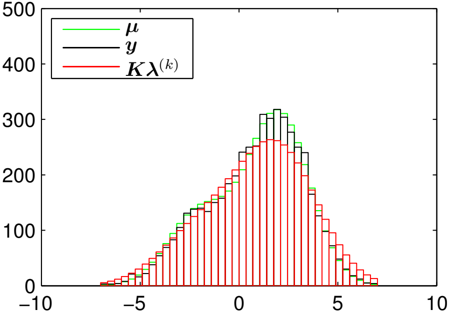


Figure: Smeared histogram

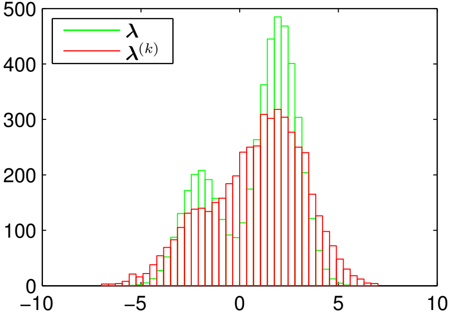


Figure: True histogram

D'Agostini demo, $k = 100$

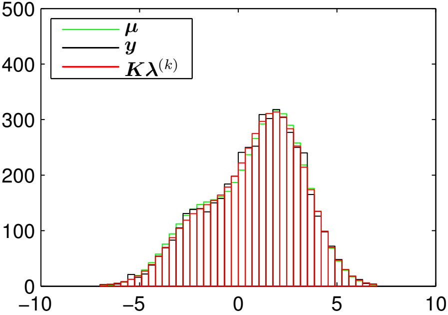


Figure: Smeared histogram

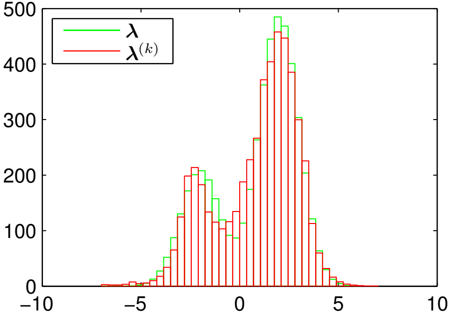


Figure: True histogram

D'Agostini demo, $k = 10000$

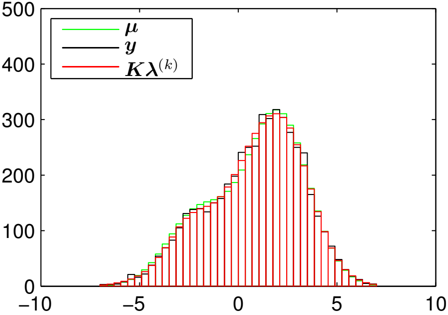


Figure: Smeared histogram

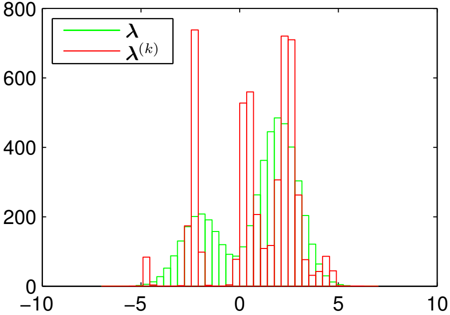


Figure: True histogram

D'Agostini demo, $k = 100000$

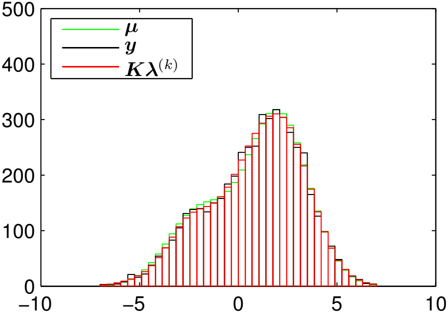


Figure: Smeared histogram

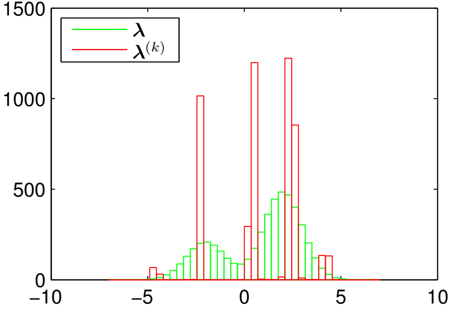
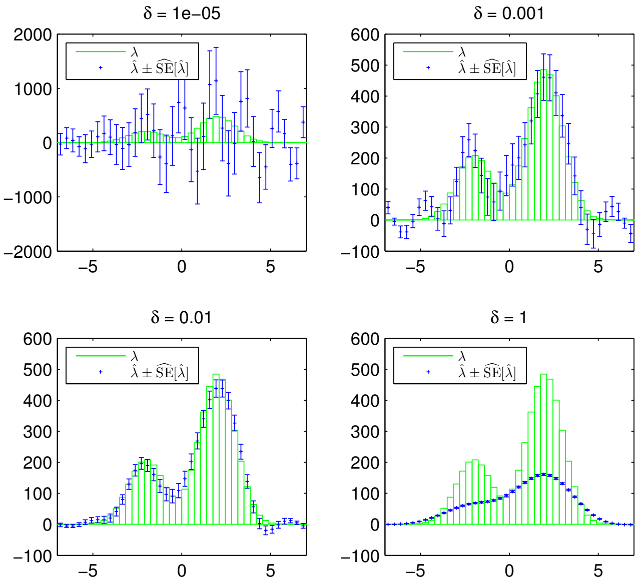


Figure: True histogram

Choice of the regularization strength

- A key issue in unfolding concerns the choice of the regularization strength (δ in Tikhonov, # of iterations in D'Agostini)
  - Used to adjust the bias-variance trade-off inherent in regularization
  - The solution and especially the uncertainties depend heavily on this choice
- This choice should ideally be done using an objective data-driven criterion
  - In particular, one must not rely on the software defaults for the regularization strength (such as 4 iterations of D'Agostini in `RooUnfold`)
- Many data-driven methods have been proposed:
  1. (Weighted/generalized) cross-validation (e.g., Green and Silverman, 1994)
  2. L-curve (Hansen, 1992)
  3. Marginal maximum likelihood (MMLE; Kuusela and Panaretos (2015))
  4. Goodness-of-fit test in the smeared space (Veklerov and Llacer, 1987)
  5. Akaike information criterion (Volobouev, 2015)
  6. Minimization of a global correlation coefficient (Schmitt, 2012)
  7. Stein's unbiased risk estimate (SURE; new in `TUnfold` V17.9)
  8. ...
- Limited experience about the relative merits of these in typical unfolding problems

Tikhonov regularization, $P(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|^2$, varying δ



Uncertainty quantification in unfolding

- For the rest of this talk, let's assume that we are interested in some linear functional $\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\lambda}$ of $\boldsymbol{\lambda}$ (or potentially some collection of functionals)
  - For example, $\theta = \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{\lambda} = i$th unfolded bin
- We will use $\hat{\theta} = \boldsymbol{h}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}$ as a natural point estimator of θ
- For uncertainty quantification, our goal is to find a random interval $\left[\underline{\theta}(\boldsymbol{y}), \overline{\theta}(\boldsymbol{y})\right]$ with *coverage probability* $1-\alpha$:

$$\mathrm{P}\left(\theta \in \left[\underline{\theta}(\boldsymbol{y}), \overline{\theta}(\boldsymbol{y})\right]\right) \approx 1-\alpha$$

- Most implementations construct the interval based on the variance of $\hat{\theta}$:

$$[\underline{\theta}, \overline{\theta}] = \left[\hat{\theta} - z_{1-\alpha/2}\sqrt{\mathrm{var}(\hat{\theta})},\, \hat{\theta} + z_{1-\alpha/2}\sqrt{\mathrm{var}(\hat{\theta})}\right]$$

- But: These intervals may suffer from significant undercoverage because they ignore the regularization bias

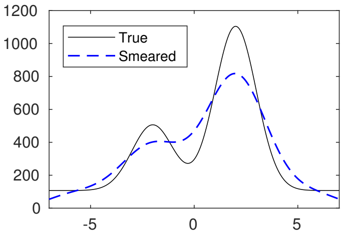
Regularization and frequentist coverage

In fact, if we approximate the Poisson noise using a Gaussian and use an affine estimator $\hat{\boldsymbol{\lambda}}$ (e.g., Tikhonov-type estimators), then the coverage of the variability intervals can be written down in closed form (Kuusela, 2016):

$$\mathbb{P}(\theta \in [\underline{\theta}, \overline{\theta}]) = \Phi\left(\frac{\mathrm{bias}(\hat{\theta})}{\sqrt{\mathrm{var}(\hat{\theta})}} + z_{1-\alpha/2}\right) - \Phi\left(\frac{\mathrm{bias}(\hat{\theta})}{\sqrt{\mathrm{var}(\hat{\theta})}} - z_{1-\alpha/2}\right)$$

The intervals have coverage $1-\alpha$ if and only if $\mathrm{bias}(\hat{\theta}) = 0$; otherwise coverage $< 1-\alpha$ and symmetric w.r.t. the sign of $\mathrm{bias}(\hat{\theta})$

Simulation setup



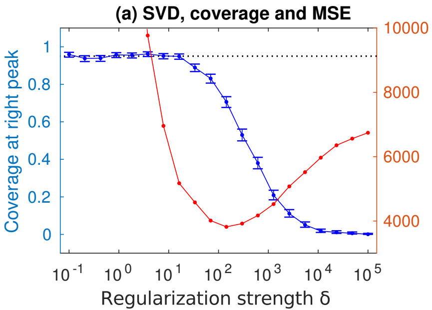
$$f(t) = \lambda_{\text{tot}} \left\{ \pi_1 \mathcal{N}(t|-2,1) + \pi_2 \mathcal{N}(t|2,1) + \pi_3 \frac{1}{|T|} \right\}$$

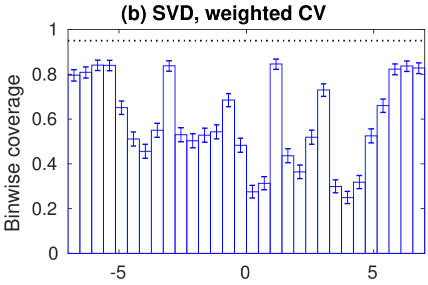
$$g(s) = \int_T \mathcal{N}(s-t|0,1) f(t)\,\mathrm{d}t$$

$$f^{\mathrm{MC}}(t) = \lambda_{\text{tot}} \left\{ \pi_1 \mathcal{N}(t|-2,1.1^2) + \pi_2 \mathcal{N}(t|2,0.9^2) + \pi_3 \frac{1}{|T|} \right\}$$

[Or slight variations of this setup.]

Undercoverage in unfolding





Coverage in SVD unfolding: as a function of the regularization strength (left) and for cross-validated regularization strength (right)

- The optimal point estimator in terms of the MSE has a sizeable regularization bias
- As a result, the unfolded variability intervals have substantial undercoverage
- Similar conclusions hold for other common methods (D'Agostini, TUnfold,...)

Wide-bin unfolding

An alternative approach that has become increasingly popular in LHC data analysis is to simply use very few unfolded bins p

$\Rightarrow$ Regularization using wide bins

Intuition: The detector should not be able to recover features smaller than its intrinsic resolution so should chose

$$\text{bin size} \gtrsim \text{detector resolution}$$

This intuition is sound but the typical implementation is problematic

Wide-bin unfolding

The response matrix elements are:

$$K_{i,j} = \frac{\int_{S_i}\int_{T_j} k(s,t) f(t)\,\mathrm{d}t\,\mathrm{d}s}{\int_{T_j} f(t)\,\mathrm{d}t}$$

This depends on the unknown intensity function f (specifically, the shape of f inside the true bins T_j)

To get around this, $K_{i,j}$ is approximated based on a MC ansatz f^{MC}:

$$K_{i,j}^{\mathrm{MC}} = \frac{\int_{S_i}\int_{T_j} k(s,t) f^{\mathrm{MC}}(t)\,\mathrm{d}t\,\mathrm{d}s}{\int_{T_j} f^{\mathrm{MC}}(t)\,\mathrm{d}t}$$

This means that unfolding is performed using an approximate matrix $\boldsymbol{K}^{\mathrm{MC}}$ instead of the true matrix $\boldsymbol{K}$

When p is small, one can typically unfold simply using the unregularized generalized least-squares estimator

$$\hat{\boldsymbol{\lambda}}^{\mathrm{MC}} = ((\boldsymbol{K}^{\mathrm{MC}})^T \boldsymbol{C}^{-1} \boldsymbol{K}^{\mathrm{MC}})^{-1} (\boldsymbol{K}^{\mathrm{MC}})^T \boldsymbol{C}^{-1} \boldsymbol{y}$$

But this is biased because $\boldsymbol{K}^{\mathrm{MC}} \neq \boldsymbol{K} \Rightarrow$ Wide-bin bias

Wide-bins-via-fine-bins unfolding

Because of the wide-bin bias, variability intervals based on $\hat{\boldsymbol{\lambda}}^{\mathrm{MC}}$ will undercover

Again, we could try to inflate the intervals by an amount corresponding to the bias, but, as before, this bias is very difficult to estimate and quantify

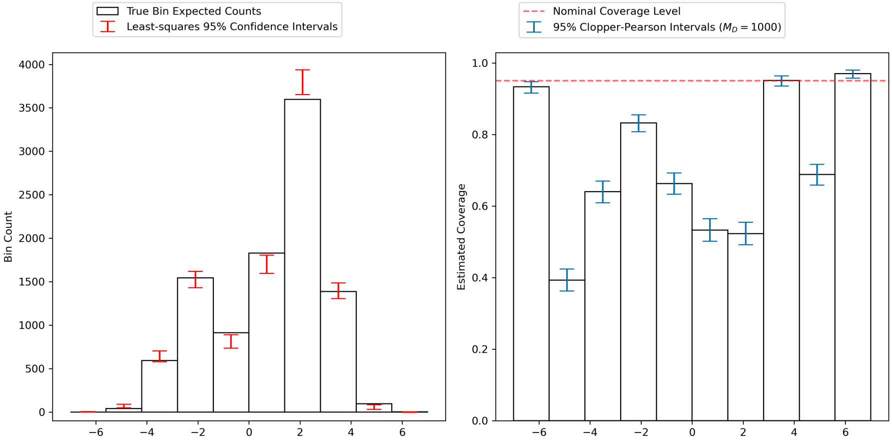
Alternative idea (Stanley et al., 2022):

The wide-bin bias gets reduced the smaller the bins in the true space

So we can *first unfold with fine bins (and no regularization) and then aggregate into wide bins, keeping track of the bin-to-bin correlation in the error propagation*

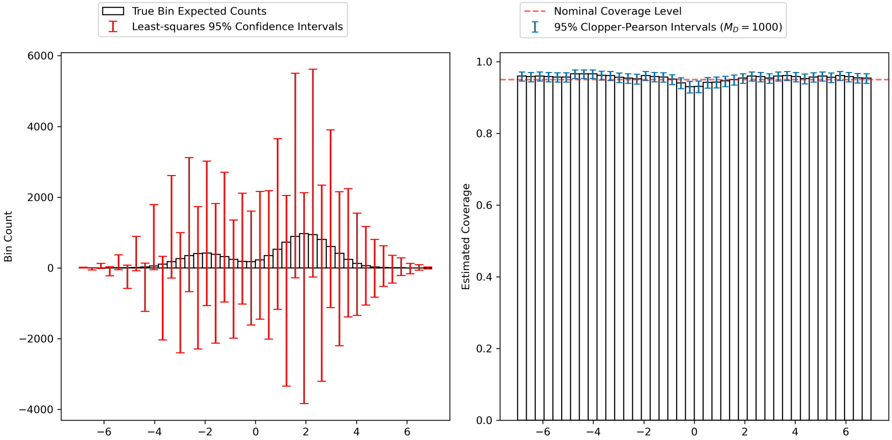
This wide-bins-via-fine-bins unfolding approach provides reasonably sized unfolded confidence intervals that do not suffer from the regularization bias and have minimal wide-bin bias

Wide bins, standard approach, misspecified MC



Intervals undercover because they ignore the wide-bin bias caused by the misspecified f^{MC}

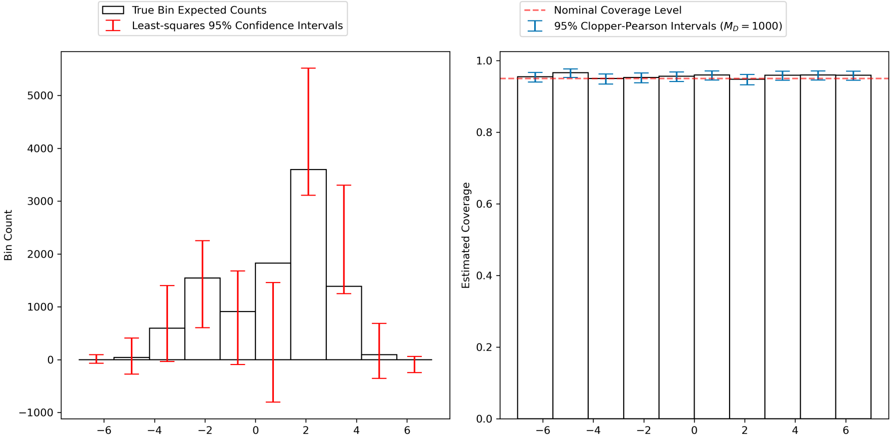
Fine bins, standard approach, misspecified MC



With narrow bins, less dependence on f^{MC} so coverage is improved, but the intervals are very wide

$\Rightarrow$ Let's aggregate these into wide bins

Wide bins via fine bins, misspecified MC



With the same misspecified f^{MC}, wide-bins-via-fine-bins unfolding gives both proper coverage and reasonably sized intervals

Handling constraints and rank-deficient matrices

The previous example shows that the wide-bins-via-fine-bins approach can circumvent both the regularization bias and the wide-bin bias

But the simple approach based on the least-squares variability intervals has two important limitations:

- It cannot easily impose constraints (such as positivity) on the solution
- It cannot handle column-rank-deficient response matrices $\boldsymbol{K}$ (such as when $\#$ of true bins $>$ $\#$ of smeared bins)

Handling constraints and rank-deficient matrices

We have recently developed[2] two new methods that can incorporate constraints and handle rank-deficient matrices while preserving coverage:

- One-at-a-time strict bounds (OSB) intervals
- Prior-optimized (PO) intervals

The OSB intervals are a modification of the simultaneous strict bounds (SSB) intervals of Stark (1992) where the intervals are calibrated to have binwise coverage instead of simultaneous coverage

The PO intervals are decision-theoretic intervals where the interval length is optimized using a prior subject to a constraint on correct coverage[3]

Both intervals have correct coverage empirically; PO also has a rigorous proof of coverage; details in Stanley et al. (2022)

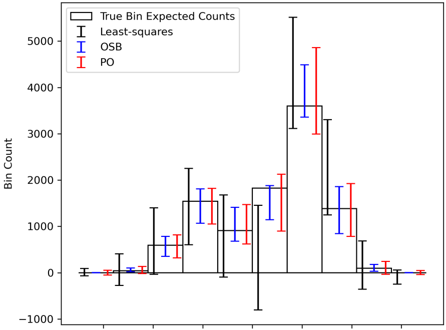
---

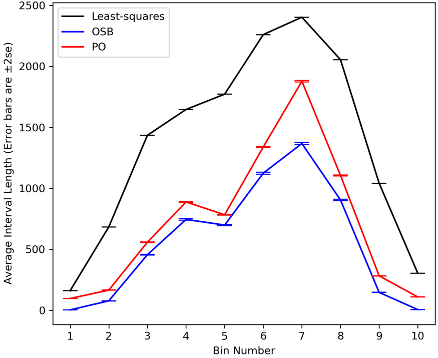
[2] M. Stanley, P. Patil, and M. Kuusela, Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals, *Journal of Instrumentation*, 17(10):P10013, 2022

[3] Importantly, finite-sample frequentist coverage is guaranteed even for misspecified priors, but the interval length might be suboptimal in those cases.

Wide bins via fine bins, with positivity constraint

The interval lengths can be reduced by imposing a positivity constraint on the solution:



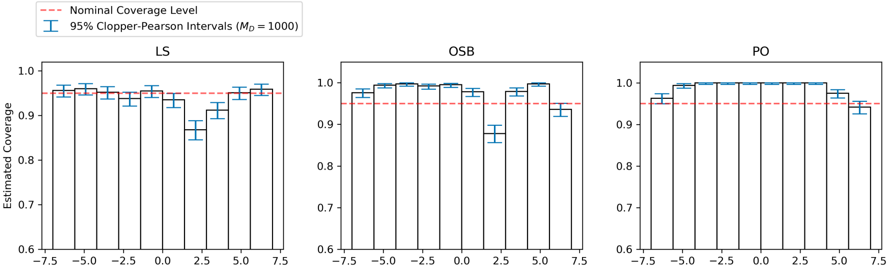


All of the above intervals have proper coverage

Motivation for the rank-deficient case

However, even with a 40×40 response matrix, the wide-bin bias can be sizeable for heavily misspecified f^{MC}

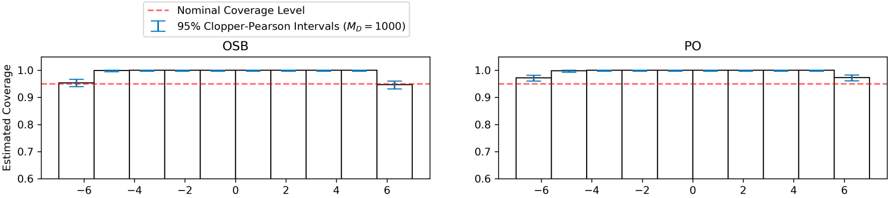
Coverage of the previous three methods for an adversarial f^{MC}:



Wide bins via fine bins, with rank-deficient $\boldsymbol{K}$

This can be fixed by using an even larger number of true bins, which requires methods that can handle a rank-deficient $\boldsymbol{K}$

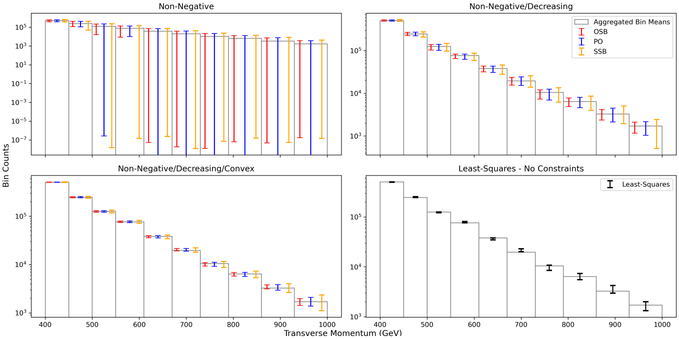
Coverage of the OSB and PO intervals with a 40×80 response matrix:



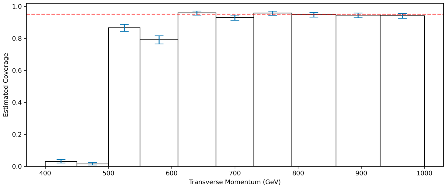
We have additionally found that:

- The interval width of both methods flattens out as the number of true bins is further increased
- The PO interval width has little sensitivity to the choice of the prior

Application to unfolding a steeply falling spectrum



The OSB, PO and SSB intervals based on a 30×60 response matrix all have at least 95% coverage, while the least-squares intervals with a 30×10 matrix do not cover:



Conclusions

- Unfolding is a complex data analysis task with many potential pitfalls
- Regularization works well for point estimation, but uncertainty quantification based on regularized estimators is very difficult
- Uncertainties derived using standard regularization methods can have drastically lower frequentist coverage than expected due to the regularization bias
- As a result, regularization-free wide-bin unfolding has become increasingly popular
- But this creates a non-trivial wide-bin bias which is equally difficult to quantify accurately
- Wide-bins-via-fine-bins unfolding provides a potential solution
  - See Stanley et al. (2022) for methods and simulation results

References I

T. Adye. Unfolding algorithms and tests using RooUnfold. In H. B. Prosper and L. Lyons, editors, *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, CERN-2011-006, pages 313–318, CERN, Geneva, Switzerland, 17–20 January 2011.

V. Blobel. Unfolding methods in high-energy physics experiments. In *CERN Yellow Report 85-09*, pages 88–127, 1985.

V. Blobel, The RUN manual: Regularized unfolding for high-energy physics experiments, OPAL Technical Note TN361, 1996.

J. Bourbeau and Z. Hampel-Arias, PyUnfold: A Python package for iterative unfolding, *The Journal of Open Source Software*, 3(26):741, 2018.

A. Bozson, G. Cowan, and F. Spanò, Unfolding with Gaussian processes, 2018.

G. Choudalakis. Fully Bayesian unfolding. arXiv:1201.4612v4 [physics.data-an], 2012.

G. D'Agostini, A multidimensional unfolding method based on Bayes' theorem, *Nuclear Instruments and Methods A*, 362:487–498, 1995.

A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

References II

P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, 1994.

P. C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Review*, 34(4):561–580, 1992.

A. Höcker and V. Kartvelishvili, SVD approach to data unfolding, *Nuclear Instruments and Methods in Physics Research A*, 372:469–481, 1996.

M. Kuusela. *Uncertainty quantification in unfolding elementary particle spectra at the Large Hadron Collider*. PhD thesis, EPFL, 2016. Available online at: `https://infoscience.epfl.ch/record/220015`.

M. Kuusela and V. M. Panaretos, Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification, *The Annals of Applied Statistics*, 9(3):1671–1705, 2015.

M. Kuusela and P. B. Stark, Shape-constrained uncertainty quantification in unfolding steeply falling elementary particle spectra, *The Annals of Applied Statistics*, 11(3): 1671–1710, 2017.

K. Lange and R. Carson, EM reconstruction algorithms for emission and transmission tomography, *Journal of Computer Assisted Tomography*, 8(2):306–316, 1984.

References III

L. B. Lucy, An iterative technique for the rectification of observed distributions, *Astronomical Journal*, 79(6):745–754, 1974.

B. Malaescu. An iterative, dynamically stabilized (IDS) method of data unfolding. In H. B. Prosper and L. Lyons, editors, *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, CERN-2011-006, pages 271–275, CERN, Geneva, Switzerland, 17–20 January 2011.

N. Milke, M. Doert, S. Klepser, D. Mazin, V. Blobel, and W. Rhode, Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics, *Nuclear Instruments and Methods in Physics Research A*, 697:133–147, 2013.

W. H. Richardson, Bayesian-based iterative method of image restoration, *Journal of the Optical Society of America*, 62(1):55–59, 1972.

S. Schmitt, TUnfold, an algorithm for correcting migration effects in high energy physics, *Journal of Instrumentation*, 7:T10003, 2012.

L. A. Shepp and Y. Vardi, Maximum likelihood reconstruction for emission tomography, *IEEE Transactions on Medical Imaging*, 1(2):113–122, 1982.

M. Stanley, P. Patil, and M. Kuusela, Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals, *Journal of Instrumentation*, 17(10):P10013, 2022.

References IV

P. B. Stark, Inference in infinite-dimensional inverse problems: Discretization and duality, *Journal of Geophysical Research*, 97(B10):14055–14082, 1992.

A. M. Stuart, Inverse problems: A Bayesian perspective, *Acta Numerica*, 19:451–559, 2010.

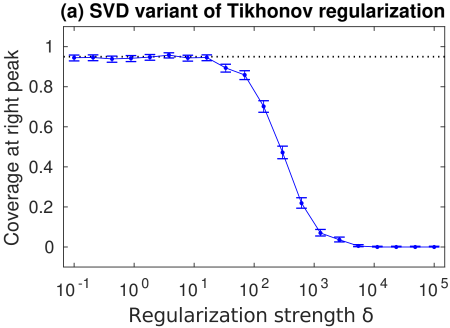
Y. Vardi, L. A. Shepp, and L. Kaufman, A statistical model for positron emission tomography, *Journal of the American Statistical Association*, 80(389):8–20, 1985.

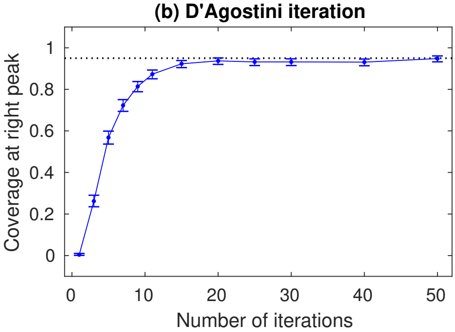
E. Veklerov and J. Llacer, Stopping rule for the MLE algorithm based on statistical hypothesis testing, *IEEE Transactions on Medical Imaging*, 6(4):313–319, 1987.

I. Volobouev. On the expectation-maximization unfolding with smoothing. arXiv:1408.6500v2 [physics.data-an], 2015.

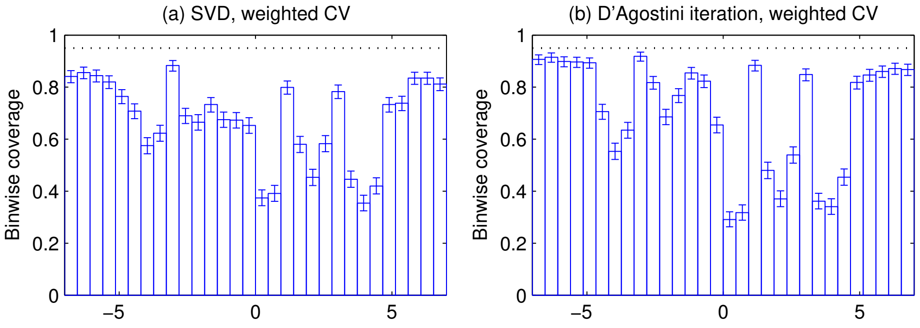
Backup

Coverage as a function of regularization strength





Undercoverage of existing methods



There is major undercoverage if regularization strength chosen using (weighted) cross-validation; same is true for L-curve and MMLE.

Key point: These methods are designed for optimal point estimation, but:

optimal point estimation $\neq$ optimal uncertainty quantification

Undersmoothed unfolding

- Standard methods for picking the regularization strength choose too much bias from the perspective of the variance-based uncertainties
- One possible solution is to *debias* the estimator, i.e., to adjust the bias-variance trade-off to the direction of less bias and more variance
- The simplest form of debiasing is to reduce δ from the cross-validation / L-curve / MMLE value until the intervals have close-to-nominal coverage
- The challenge is to come up with a data-driven rule for deciding *how much to undersmooth*
- With Lyle Kim, we have implemented the data-driven methods from Kuusela (2016) as an extension of `TUnfold`
- The code is available at:

  `https://github.com/lylejkim/UndersmoothedUnfolding`

- If you're already working with `TUnfold`, then trying this approach requires adding only one extra line of code to your analysis

Unfolded histograms, $\boldsymbol{\lambda}^{\mathrm{MC}} = 0$

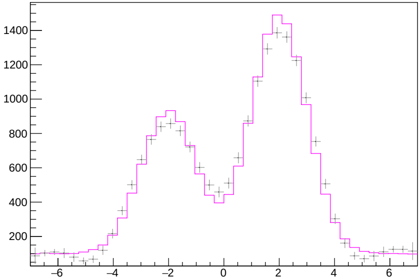


Figure: L-curve, $\tau = \sqrt{\delta}$=0.01186

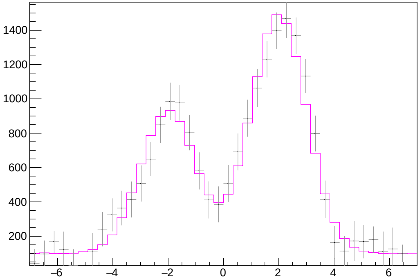


Figure: Undersmoothing, $\tau = \sqrt{\delta}$=0.00177

Binwise coverage, $\lambda^{\mathrm{MC}} = 0$

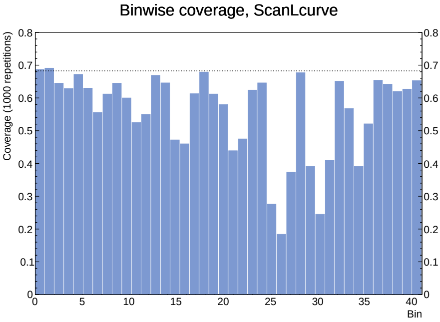


Figure: L-curve

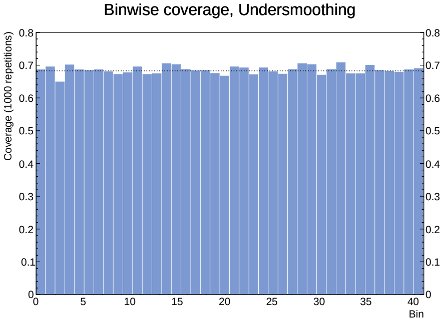


Figure: Undersmoothing

Unregularized unfolding?

- At the end of the day, *any regularization technique makes unverifiable assumptions about the true solution*
  - If these assumptions are not satisfied, the uncertainties will be wrong
  - In the absence of oracle information about the true $\boldsymbol{\lambda}$, there does not seem to be any obvious way around this
- So maybe we should reconsider whether explicit regularization is such a good idea to start with?
- Instead of finding a regularized estimator of $\boldsymbol{\lambda}$, what if we simply used[4] the unregularized matrix inverse $\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$?
- This is unbiased ($\mathrm{E}(\hat{\boldsymbol{\lambda}}) = \boldsymbol{\lambda}$) and hence also the corresponding estimator $\hat{\theta} = \boldsymbol{h}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}$ of the functional $\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\lambda}$ is unbiased
- Therefore, by the previous discussion, the resulting variability intervals have correct coverage $1 - \alpha$

[4] For simplicity, I assume here that $\boldsymbol{K} \in \mathbb{R}^{n \times p}$ is an invertible square matrix. The case where $n > p$ with $\boldsymbol{K}$ having full column rank is also easy using the pseudoinverse $\hat{\boldsymbol{\lambda}} = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y}$. The case where $\boldsymbol{K}$ is column-rank deficient (including when $p > n$) is trickier but probably doable; see https://indico.cern.ch/event/882374/.

Implicit regularization

- Of course, when $\boldsymbol{K}$ is ill-conditioned, the unregularized estimator $\hat{\boldsymbol{\lambda}}$ will have a huge variance
- *But this does not mean that $\hat{\theta} = \boldsymbol{h}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}$ needs to have a huge variance!*
- The mapping $\hat{\boldsymbol{\lambda}} \mapsto \hat{\theta} = \boldsymbol{h}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}$ can act as an implicit regularizer resulting in a well-constrained interval $[\underline{\theta}, \overline{\theta}]$ for the functional $\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\lambda}$
- This is especially the case when the functional is a smoothing / averaging / aggregation operation
  - For example, inference for aggregated unfolded bins (demo to follow)
- Of course, there are also functionals that are more difficult to constrain (e.g., individual bins $\theta = \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{\lambda}$, derivatives,...)
- In those cases, the intervals $[\underline{\theta}, \overline{\theta}]$ are wide—as they should be, since there is simply not enough information in the data $\boldsymbol{y}$ to constrain these functionals

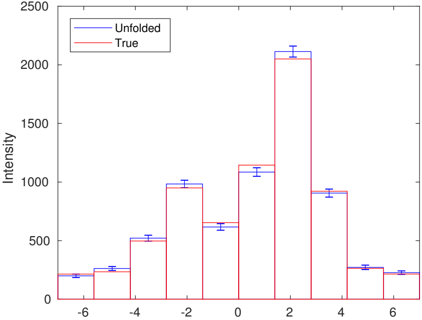
Wide bin unfolding

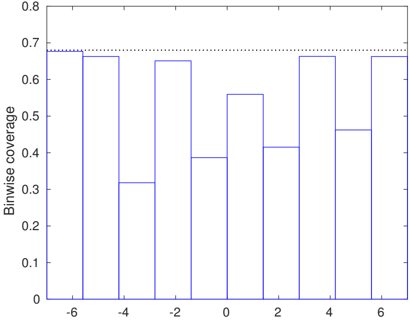
- One functional we should be able to recover without explicit regularization is the integral of f over a *wide* unfolded bin:

$$H_j[f] = \int_{T_j} f(t)\,\mathrm{d}t, \quad \text{width of } T_j \text{ large}$$

- But one cannot simply arbitrarily increase the particle-level bin size in the conventional approaches, since this increases the MC dependence of $\boldsymbol{K}$
- To circumvent this, *it is possible to first unfold with fine bins (without regularization) and then aggregate into wide bins*
- Let's see how this works using a similar deconvolution setup as before

Wide bins, standard approach, perturbed MC

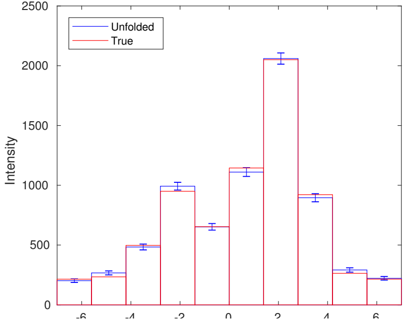


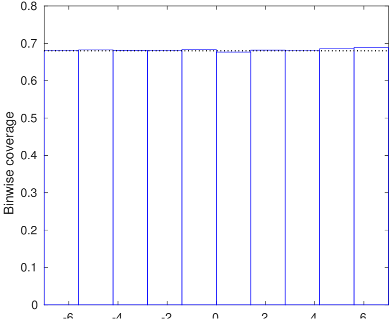


The response matrix $K_{i,j} = \frac{\int_{S_i} \int_{T_j} k(s,t) f^{\mathrm{MC}}(t)\, \mathrm{d}t\, \mathrm{d}s}{\int_{T_j} f^{\mathrm{MC}}(t)\, \mathrm{d}t}$ depends on f^{MC}

$\Rightarrow$ Undercoverage if $f^{\mathrm{MC}} \neq f$

Wide bins, standard approach, correct MC

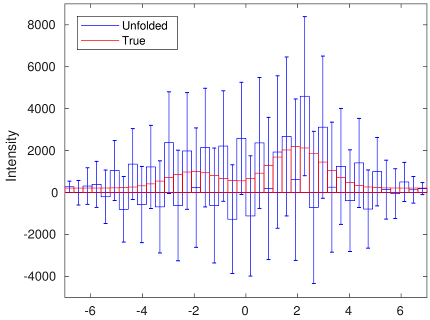


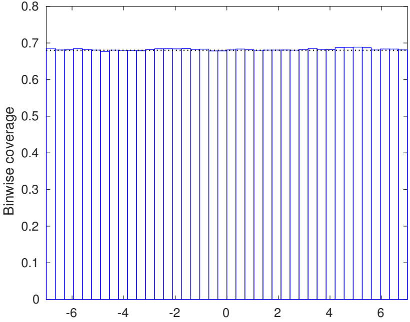


If $f^{\mathrm{MC}} = f$, coverage is correct

$\Rightarrow$ But this situation is unrealistic because f of course is unknown

Fine bins, standard approach, perturbed MC



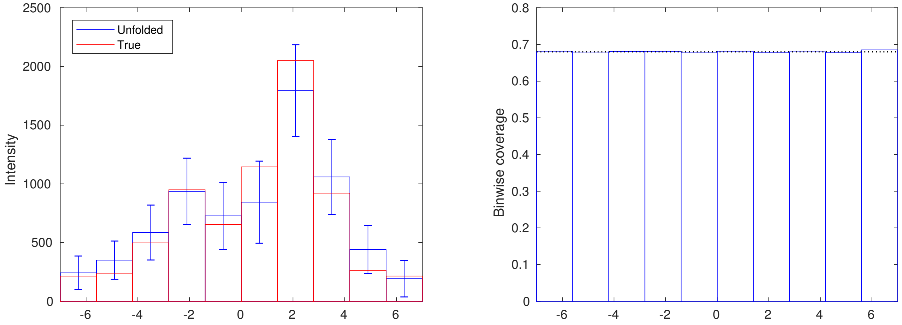


With narrow bins, less dependence on f^{MC} so coverage is correct, but the intervals are very wide[5]

$\Rightarrow$ Let's aggregate these into wide bins, keeping track of the bin-to-bin correlations in the error propagation

[5]More unfolded realizations given in the backup.

Wide bins via fine bins, perturbed MC



Wide bins via fine bins gives both correct coverage and intervals with reasonable length[6]

[6]More unfolded realizations given in the backup.

Current unfolding methods

- Two main approaches:
  1. Tikhonov regularization (i.e., SVD by Höcker and Kartvelishvili (1996) and TUnfold by Schmitt (2012)):
$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\lambda})^{\mathrm{T}} \hat{\boldsymbol{C}}^{-1} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\lambda}) + \delta P(\boldsymbol{\lambda})$$
     with
$$P_{\mathrm{SVD}}(\boldsymbol{\lambda}) = \left\| \boldsymbol{L} \begin{bmatrix} \lambda_1/\lambda_1^{\mathrm{MC}} \\ \lambda_2/\lambda_2^{\mathrm{MC}} \\ \vdots \\ \lambda_p/\lambda_p^{\mathrm{MC}} \end{bmatrix} \right\|^2 \quad \text{or} \quad P_{\mathrm{TUnfold}}(\boldsymbol{\lambda}) = \|\boldsymbol{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\mathrm{MC}})\|^2,$$
     where $\boldsymbol{L}$ is usually the discretized second derivative (also other choices possible)
  2. Expectation-maximization iteration with early stopping (D'Agostini, 1995):
$$\lambda_j^{(t+1)} = \frac{\lambda_j^{(t)}}{\sum_{i=1}^n K_{i,j}} \sum_{i=1}^n \frac{K_{i,j} y_i}{\sum_{k=1}^p K_{i,k} \lambda_k^{(t)}}, \quad \text{with } \boldsymbol{\lambda}^{(0)} = \boldsymbol{\lambda}^{\mathrm{MC}}$$
- All these methods typically regularize by biasing towards a MC ansatz $\boldsymbol{\lambda}^{\mathrm{MC}}$
- Regularization strength controlled by the choice of δ in Tikhonov or by the number of iterations in D'Agostini
- Uncertainty quantification: $[\underline{\lambda}_i, \overline{\lambda}_i] = \left[\hat{\lambda}_i - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}\left(\hat{\lambda}_i\right)}, \hat{\lambda}_i + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}\left(\hat{\lambda}_i\right)}\right]$, with $\widehat{\mathrm{var}}\left(\hat{\lambda}_i\right)$ estimated using error propagation or resampling

Coverage as a function of $\tau = \sqrt{\delta}$

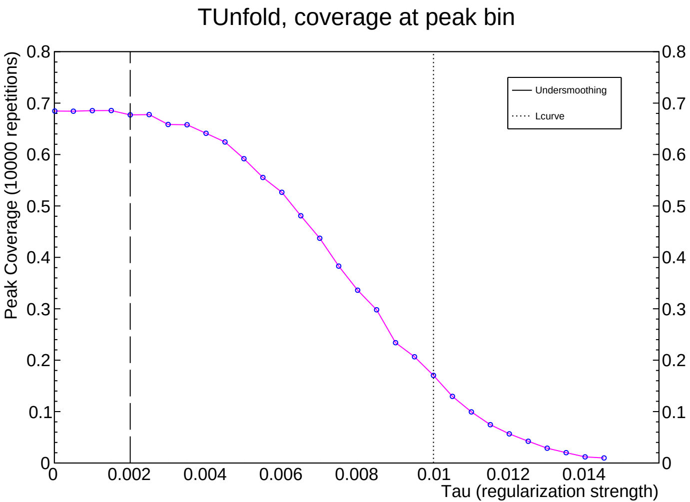
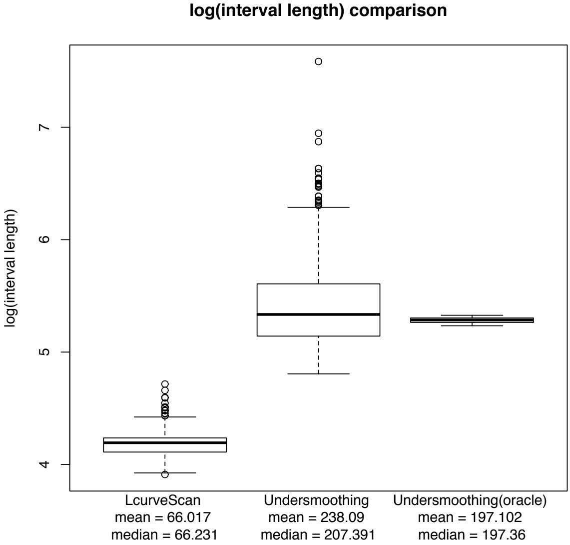
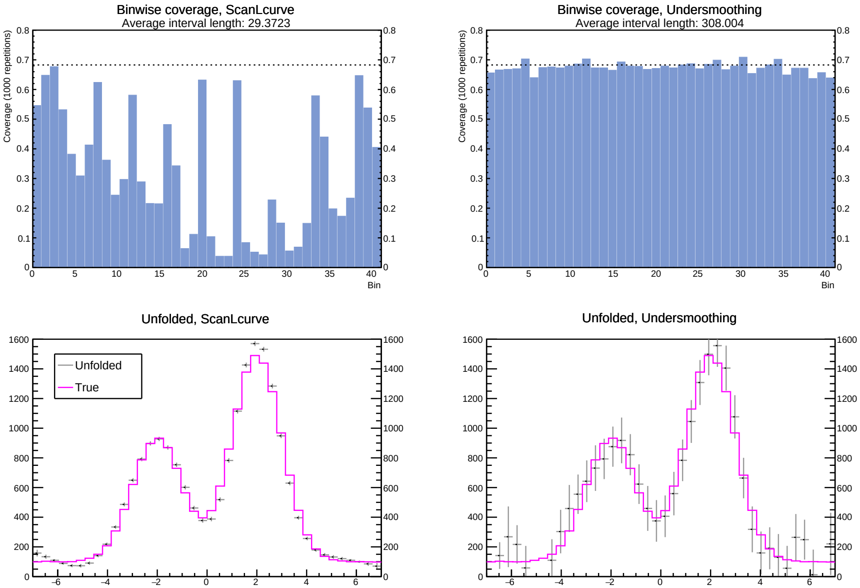


Figure: Coverage at the right peak of a bimodal density

Interval lengths, $\lambda^{\mathrm{MC}} = 0$



Histograms, coverage and interval lengths when $\lambda^{\mathrm{MC}} \neq 0$



Coverage study from Kuusela (2016)

Method	Coverage at $t = 0$	Mean length
BC (data)	0.932 *(0.915, 0.947)*	0.079 *(0.077, 0.081)*
BC (oracle)	0.937 *(0.920, 0.951)*	0.064 *(0.064, 0.064)*
US (data)	0.933 *(0.916, 0.948)*	0.091 *(0.087, 0.095)*
US (oracle)	0.949 *(0.933, 0.962)*	0.070 *(0.070, 0.070)*
MMLE	0.478 *(0.447, 0.509)*	0.030 *(0.030, 0.030)*
MISE	0.359 *(0.329, 0.390)*	0.028
Unregularized	0.952 *(0.937, 0.964)*	40316

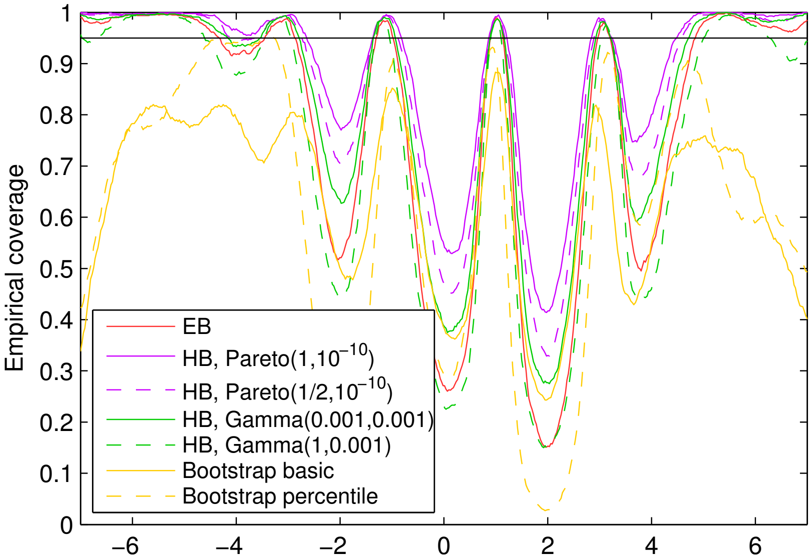
BC = iterative bias-correction

US = undersmoothing

MMLE = choose δ to maximize the marginal likelihood

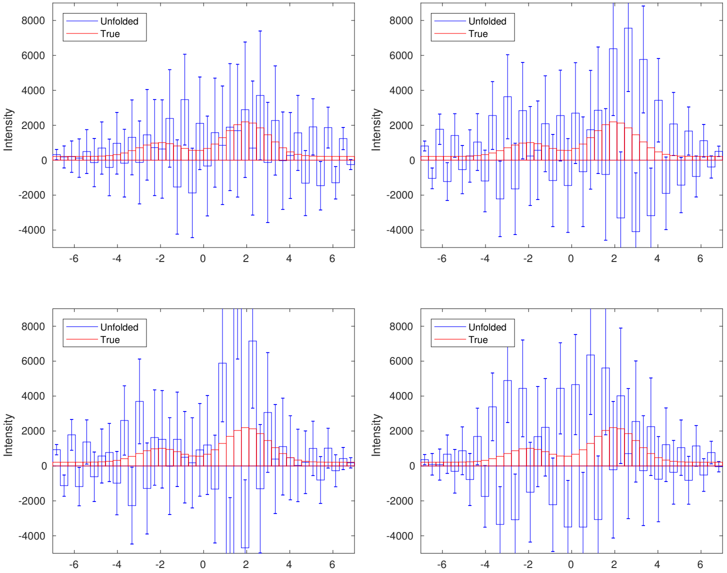
MISE = choose δ to minimize the mean integrated squared error

UQ in inverse problems is challenging



[Kuusela and Panaretos (2015)]

Fine bins, standard approach, perturbed MC, 4 realizations



Wide bins via fine bins, perturbed MC, 4 realizations