Systematic Uncertainties in Unfolding Particle Spectra

Mikael Kuusela

Department of Statistics and Data Science
Carnegie Mellon University

PHYSTAT-Systematics 2021

November 3, 2021

The unfolding problem

- Any differential cross section measurement is affected by the finite resolution of the particle detector
  - 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

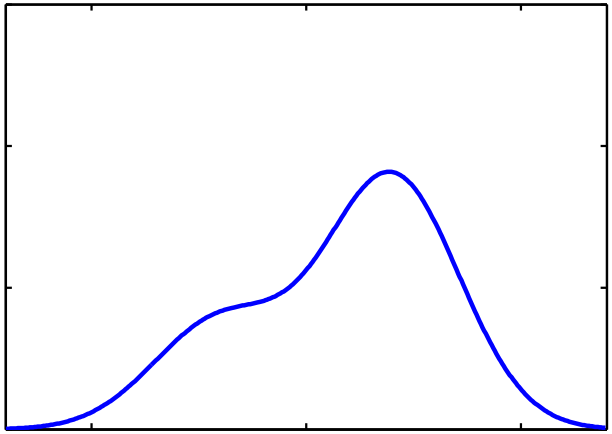
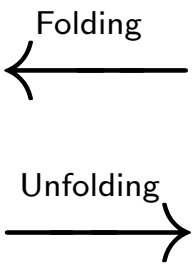


Figure: Smeared spectrum



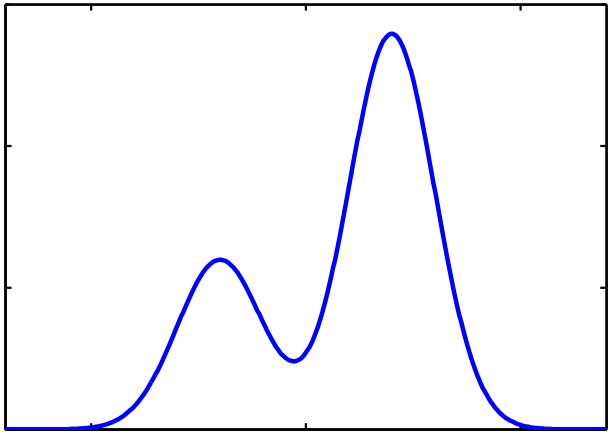


Figure: True spectrum

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

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

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

Systematic uncertainties in unfolding

Sources of systematics in unfolding:

1. Regularization bias
2. Wide-bin bias
3. Missing nuisance variables
4. Uncertainty in the response kernel k

Regularized unfolding

- When the number of true bins p is large, the response matrix $\boldsymbol{K}$ is severely ill-conditioned
- The unfolded histogram $\boldsymbol{\lambda}$ is therefore typically estimated using a regularized estimator
  - Main idea: bias $\uparrow$, variance $\downarrow$ $\Rightarrow$ $\mathrm{MSE}\downarrow$
- Two main approaches:

1. Tikhonov regularization (e.g., SVD by Höcker and Kartvelishvili (1996) and TUnfold by Schmitt (2012)):

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

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 (other choices also possible)

2. Expectation-maximization iteration with early stopping (D'Agostini, 1995):

$$\lambda_j^{(t+1)} = \frac{\lambda_j^{(t)}}{\sum_{i=1}^n K_{i,j}} \sum_{i=1}^n \frac{K_{i,j}y_i}{\sum_{k=1}^p K_{i,k}\lambda_k^{(t)}}, \text{ with } \boldsymbol{\lambda}^{(0)} = \boldsymbol{\lambda}^{\mathrm{MC}}$$

- These methods typically regularize by creating a bias toward a MC ansatz $\boldsymbol{\lambda}^{\mathrm{MC}}$

Uncertainty quantification in regularized unfolding

- Assume that we are interested in some linear functional $\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\lambda}$ of $\boldsymbol{\lambda}$ (or 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* $1-\alpha$:

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

- Most implementations construct this 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:** When $\hat{\boldsymbol{\lambda}}$ is a regularized estimator, these intervals may suffer from significant undercoverage because they ignore the regularization bias

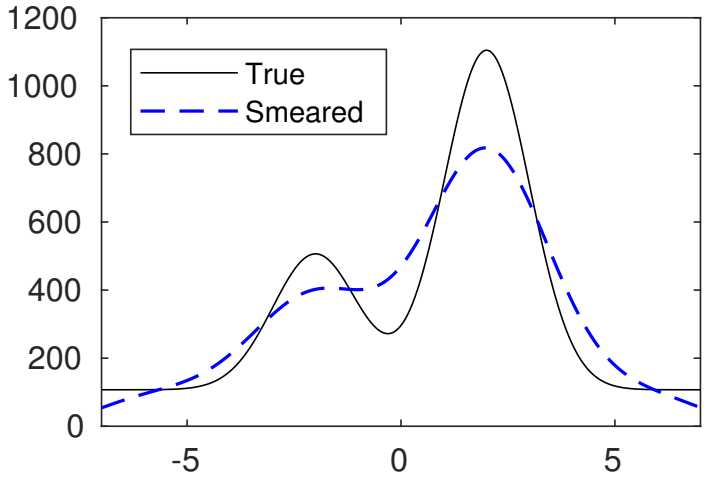
Regularization and frequentist coverage

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 $\mathrm{bias}(\hat{\theta})$

Unfolding: Simulation setup



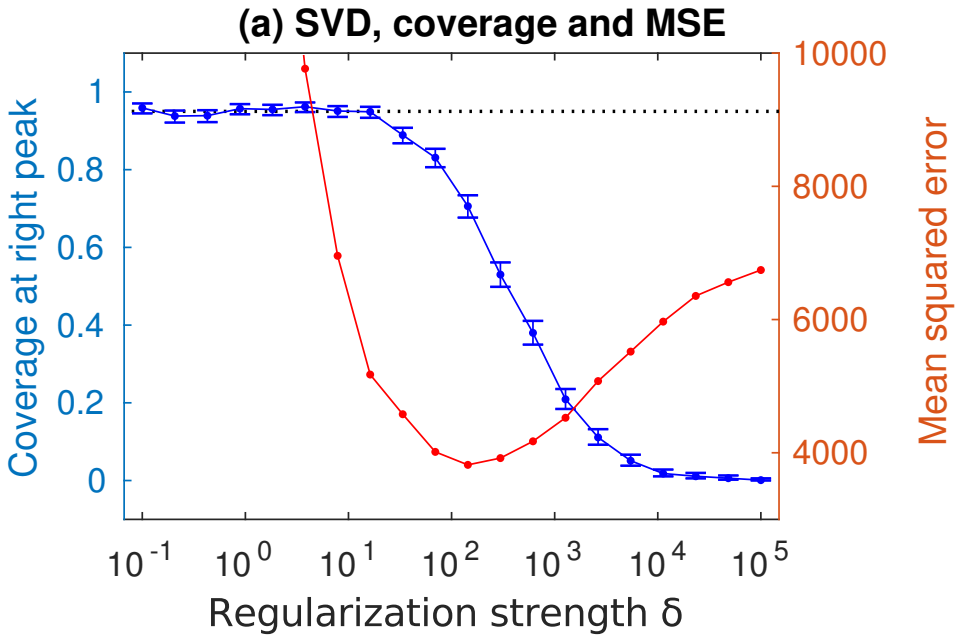
$$f(t) = \lambda_{\mathrm{tot}} \left\{ \pi_1 \mathcal{N}(t|-2,1) + \pi_2 \mathcal{N}(t|2,1) + \pi_3 \frac{1}{|T|} \right\}$$

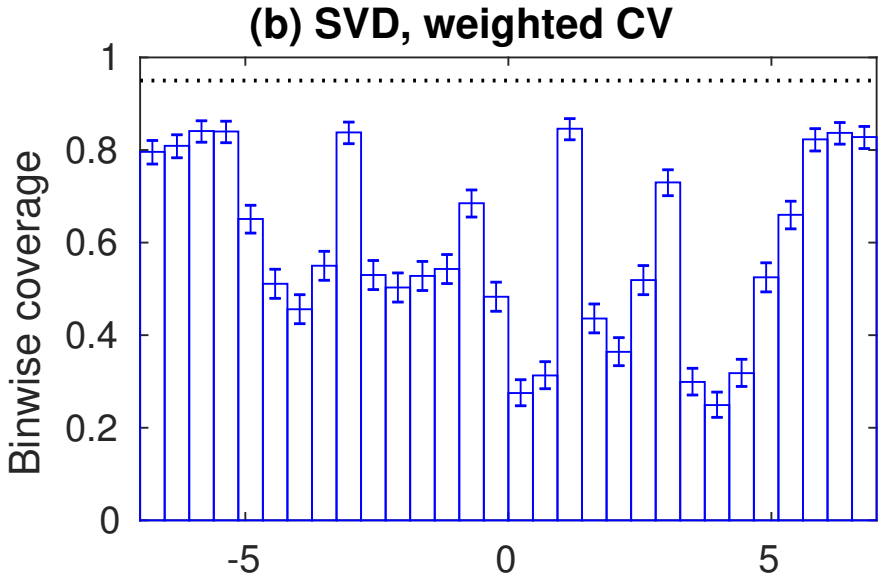
$$g(s) = \int_T \mathcal{N}(s-t|0,1) f(t)\,\mathrm{d}t$$

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

[Or slight variations of this setup.]

Undercoverage in unfolding

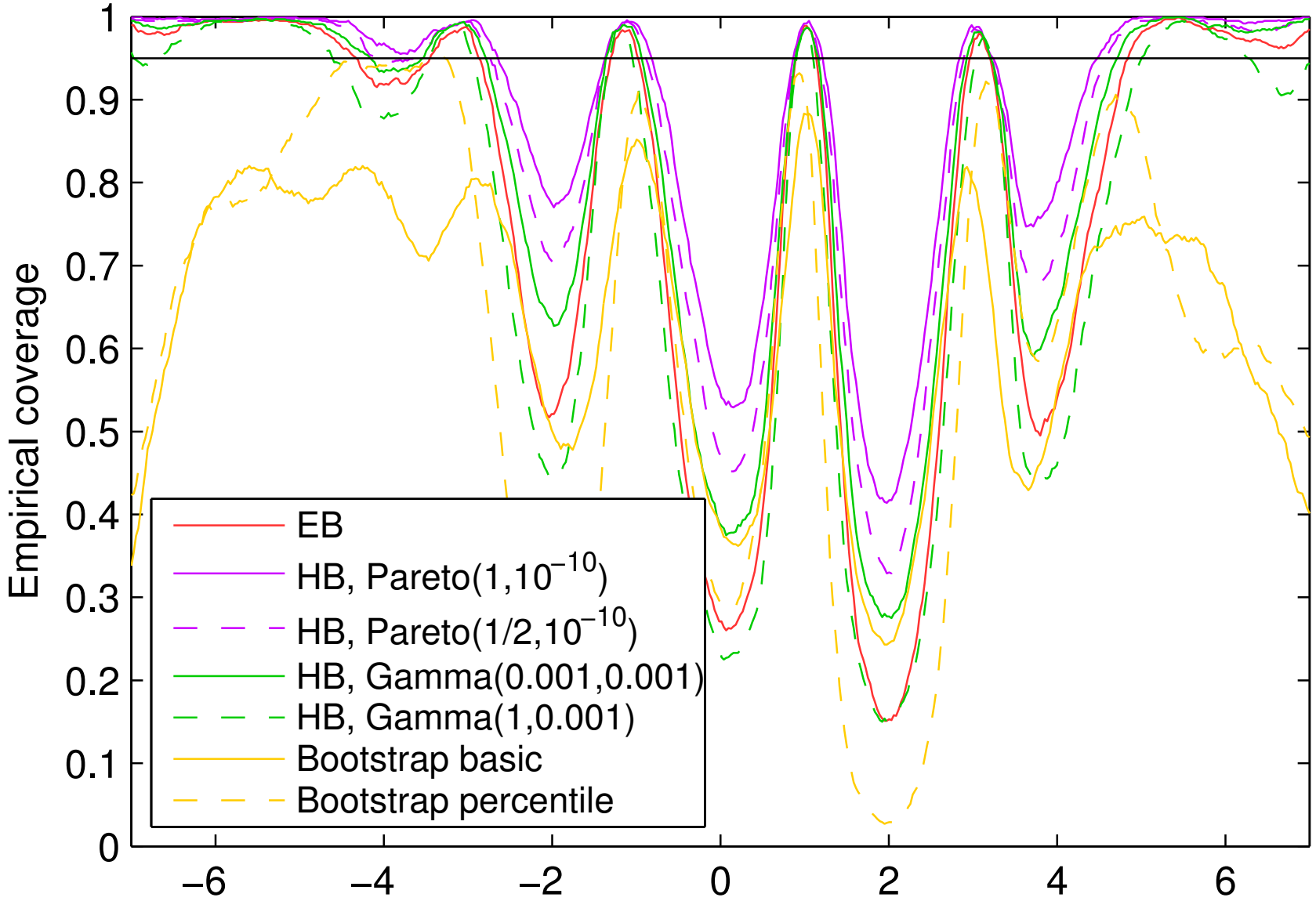




Coverage in SVD unfolding: as a function of the regularization strength (left) and for cross-validated regularization strength (right)

- 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,...)

Coverage for various unfolded confidence intervals



[Kuusela and Panaretos (2015)]

Systematic 1: Regularization bias

The regularization bias causes unfolded confidence intervals based on statistical variability to underestimate the true uncertainty

Ideally this bias should be treated as a systematic uncertainty which could then be used to inflate the confidence interval

Unfortunately, estimating the regularization bias is extremely difficult

Alternatively, it is possible to de-bias using undersmoothing (Kuusela, 2016) or iterative bias-corrections (Kuusela and Panaretos, 2015; Kuusela, 2016)

But in practice, these methods either have non-trivial breaking points or end up recovering the unregularized solution

In fact, there are good reasons to believe that it is not possible to obtain guaranteed uncertainty quantification with classical regularization (Genovese and Wasserman, 2008)

Systematic 2: Wide-bin bias

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 current 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}:

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

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., 2021):

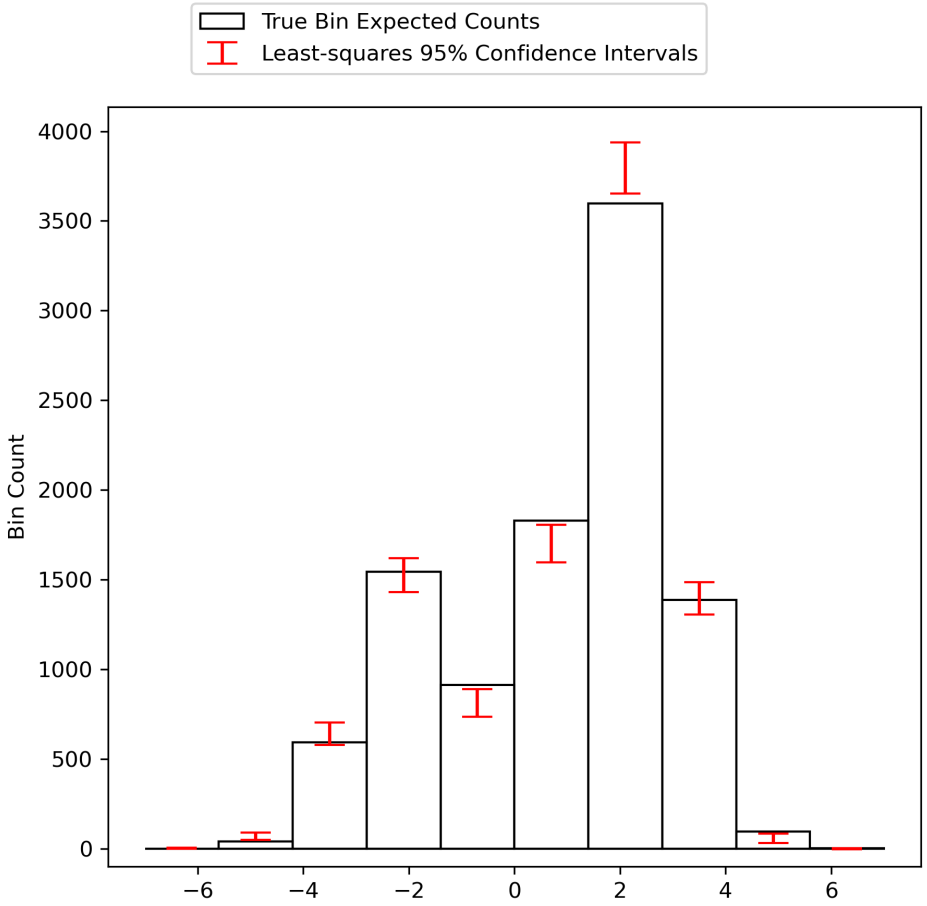
The wide-bin bias gets reduced the smaller the bins in the true space

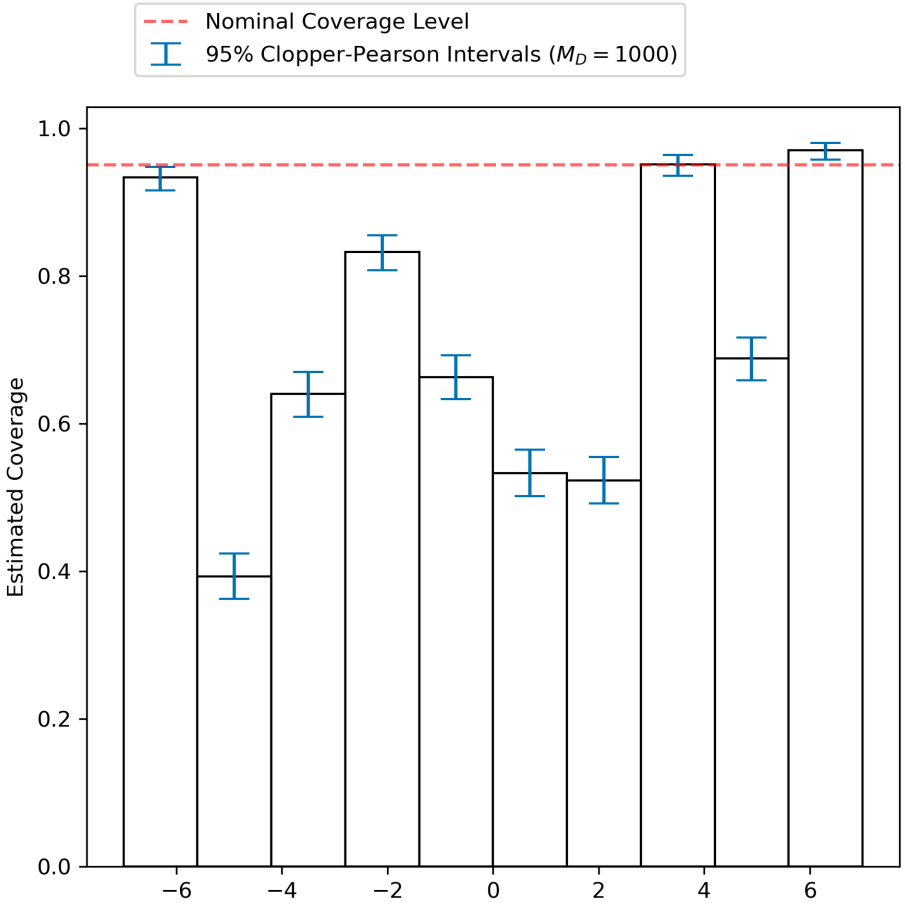
So we can *first unfold with fine bins (and no regularization) and then aggregate into wide bins, keeping track of the bin-to-bin correlation in the error propagation*

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

The following results that demonstrate this approach are joint work with Michael Stanley and Pratik Patil (both at CMU)

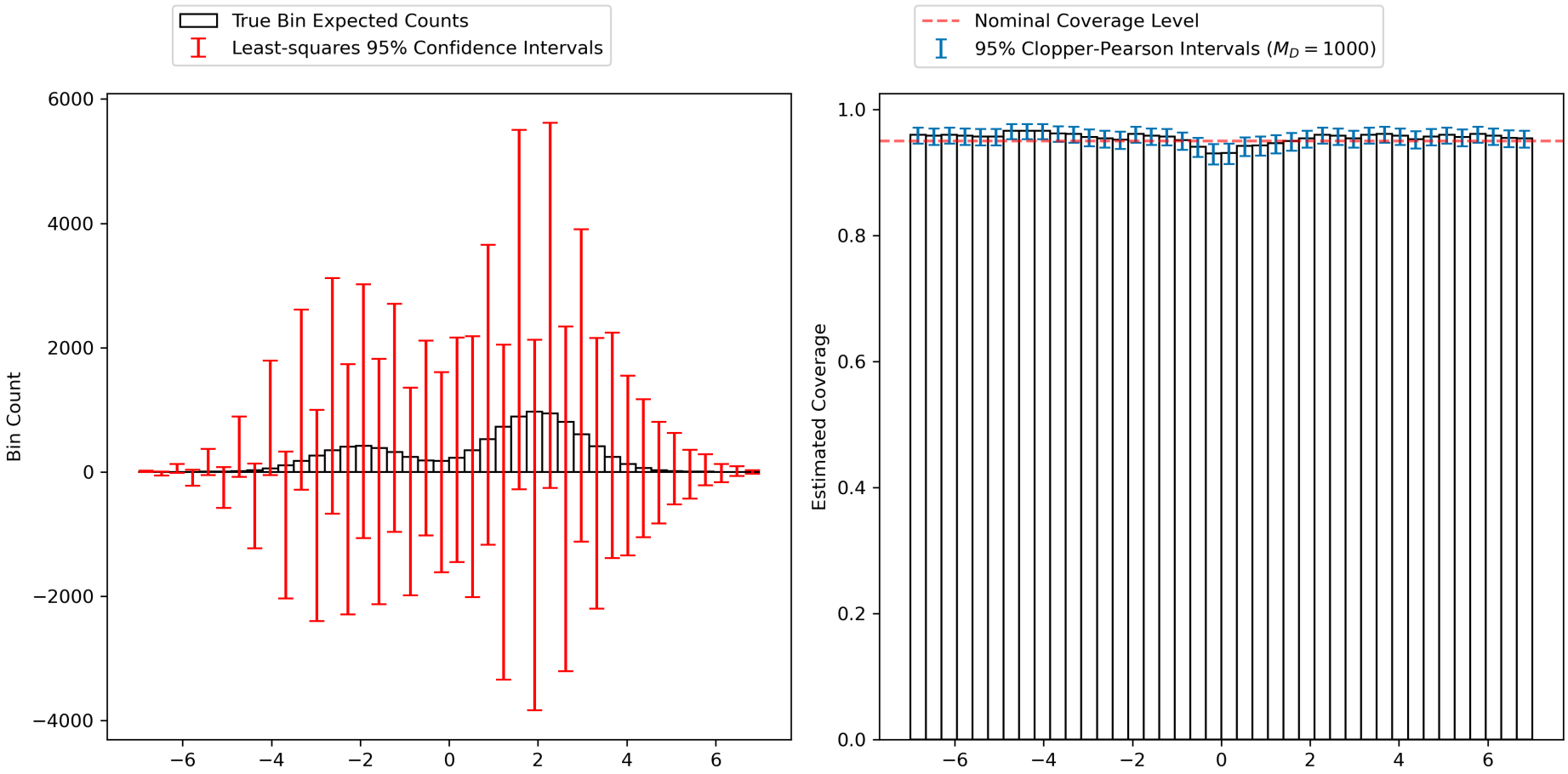
Wide bins, standard approach, misspecified MC





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

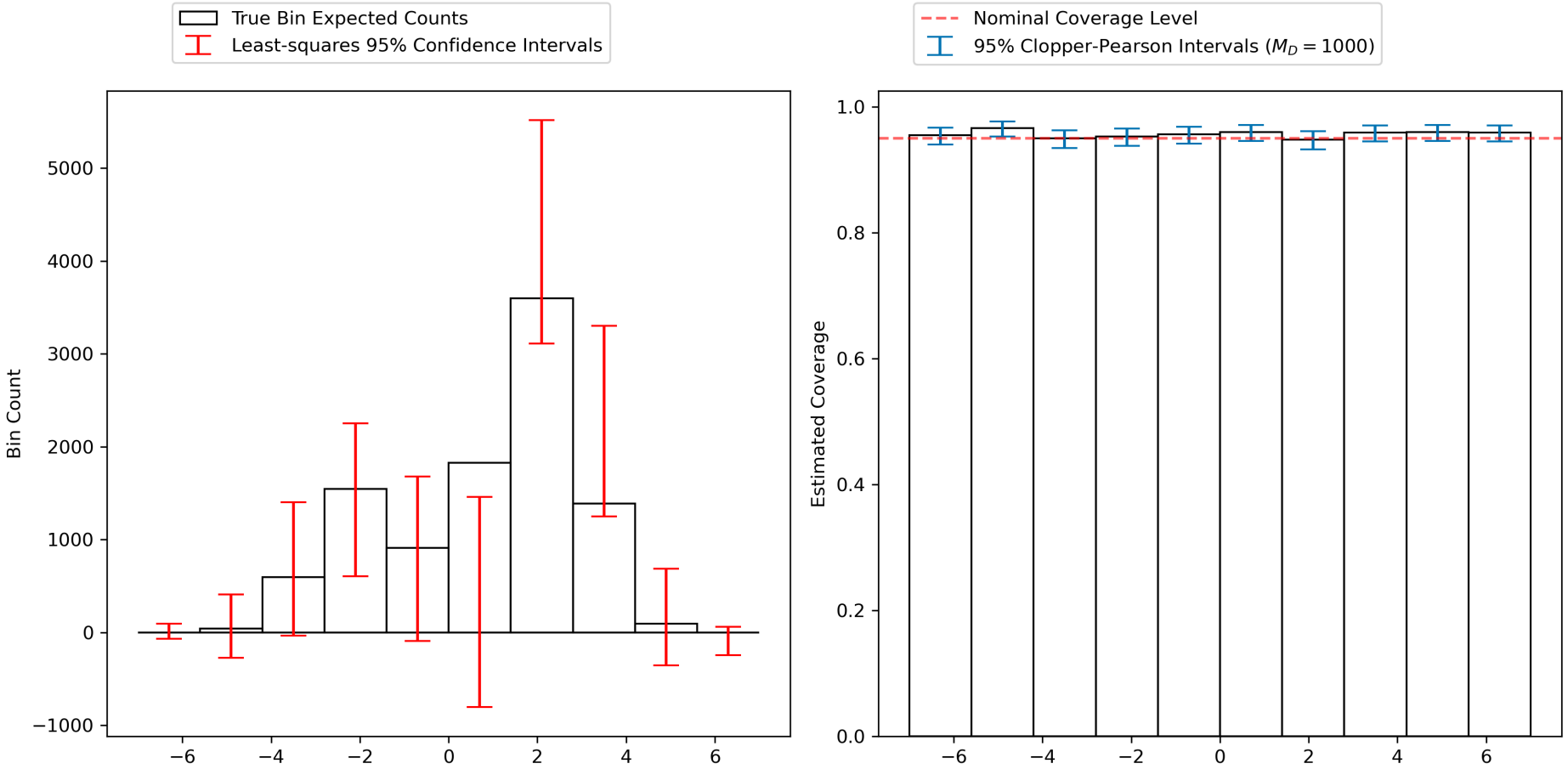
Fine bins, standard approach, misspecified MC



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[1] 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[2]

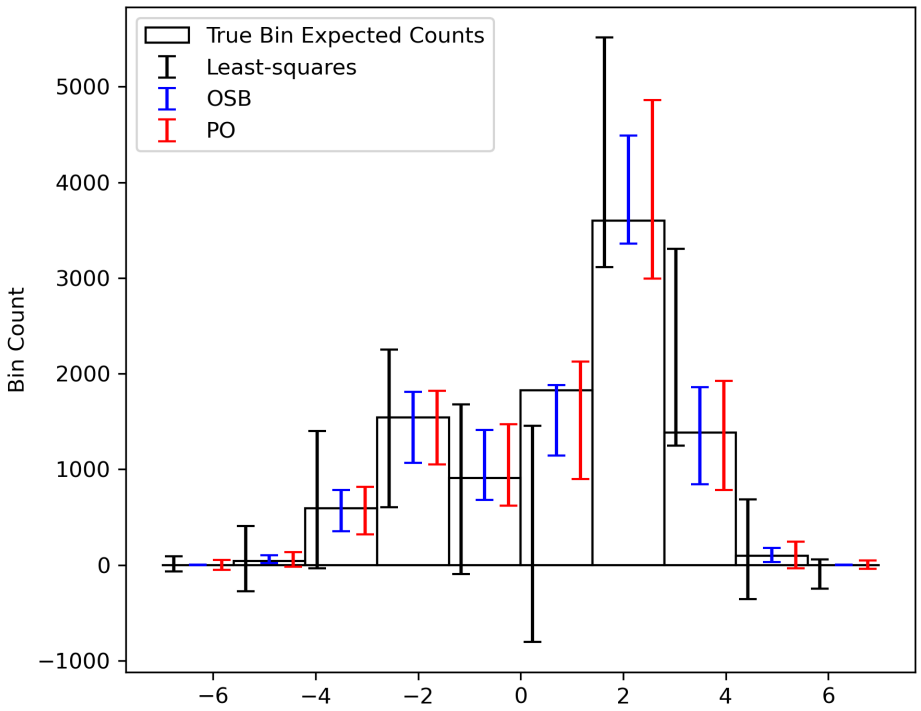
Both intervals have correct coverage empirically; PO also has a rigorous proof of coverage; details in arXiv:2111.01091

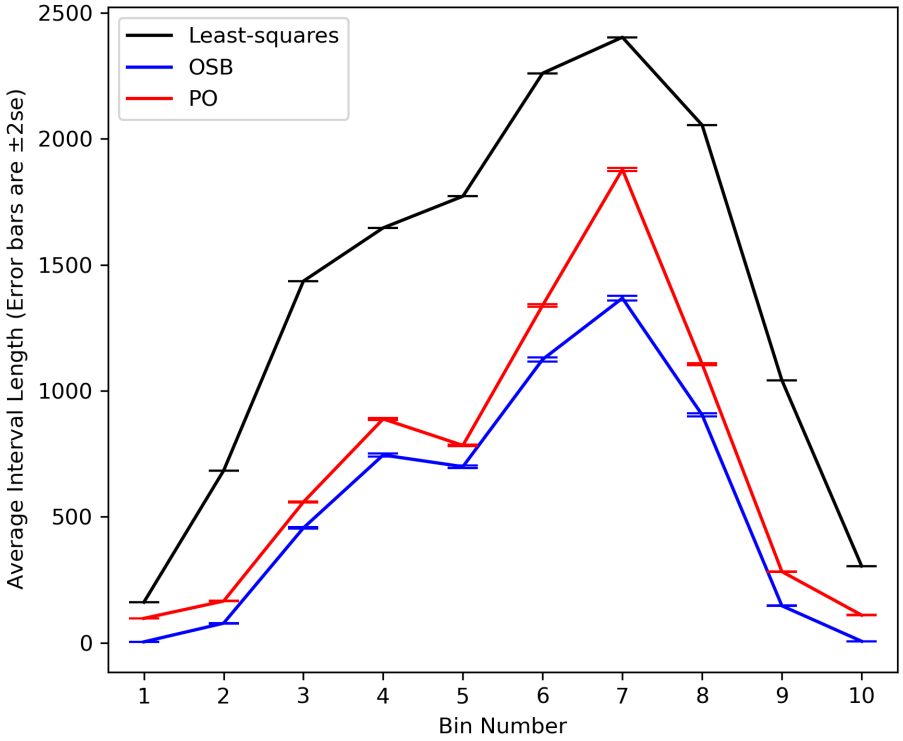
[1] M. Stanley, P. Patil, and M. Kuusela, Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals, arXiv:2111.01091 [stat.AP], 2021.

[2] 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:



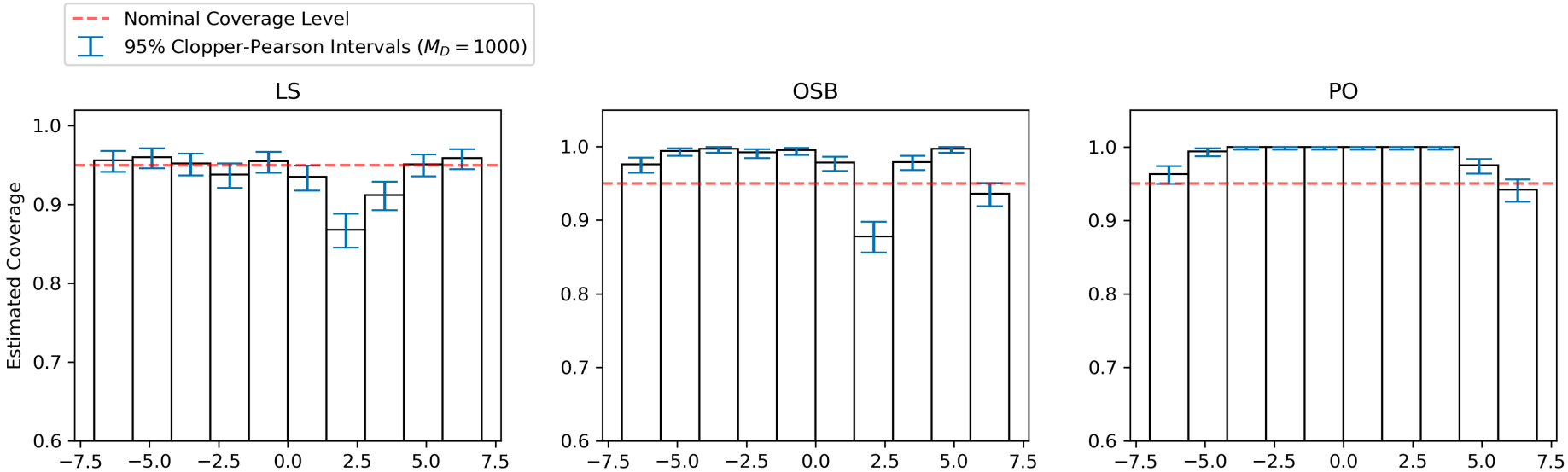


All of the above intervals have proper coverage

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}

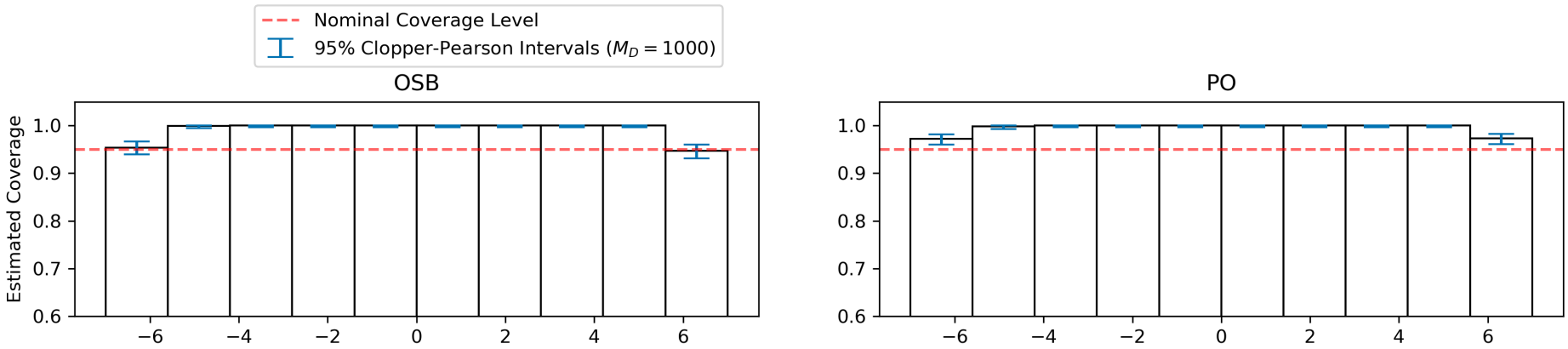
Coverage of the previous three methods for an adversarial f^{MC}:



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 than can handle a rank-deficient $\boldsymbol{K}$

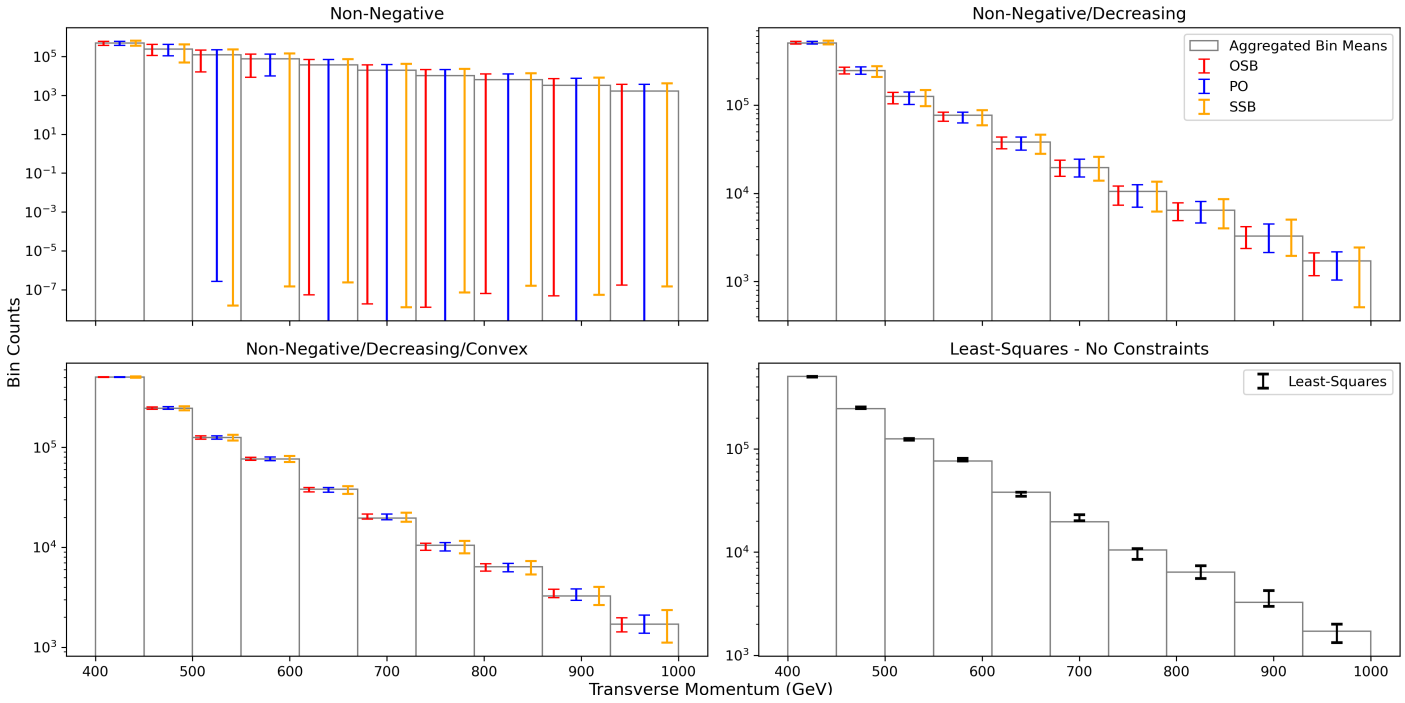
Coverage of the OSB and PO intervals with a 40×80 response matrix:



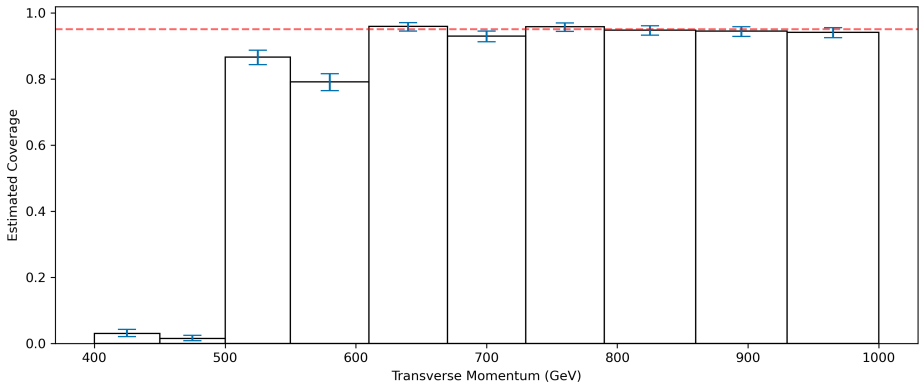
We have additionally found that:

- 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:



Systematic 3: Missing nuisance variables

In addition to the variable of interest, the detector response might depend on one or more nuisance variables

- Classical example: Energy resolution depends on pseudorapidity

Any known such nuisance variables should be included in the forward model and handled using multivariate unfolding

Let X_1 and X_2 be a variable of interest and a nuisance variable at the particle level; denote by Y_1 and Y_2 the corresponding detector-level variables

Then the appropriate forward model is

$$g(s_1, s_2) = \int\int k(s_1, s_2, t_1, t_2) f(t_1, t_2)\, \mathrm{d}t_1\, \mathrm{d}t_2,$$

where f and g are the bivariate true and smeared intensity functions and

$$k(s_1, s_2, t_1, t_2) = p(Y_1 = s_1, Y_2 = s_2 | X_1 = t_1, X_2 = t_2)$$

is the bivariate smearing kernel (ignoring efficiency, for simplicity)

But what to do about unknown nuisance variables?

Systematic 4: Uncertainty in the response kernel k

Finally, it is often the case that there is an uncertainty associated with the continuous response kernel $k(s, t)$

The kernel is typically estimated using a detector simulator

But different simulators (or configurations of those simulators) may yield different kernels

Assume that we have two simulators yielding kernels k_1 and k_2

Then one typically constructs two unfolded solutions $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_2$ corresponding to the two kernels and takes $\pm(\hat{\boldsymbol{\lambda}}_1 - \hat{\boldsymbol{\lambda}}_2)$ as a measure of systematic uncertainty

The hope is that this variation would be sufficient to contain the solution $\hat{\boldsymbol{\lambda}}_\star$ corresponding to the true kernel $k_\star$

This seems like an area where further methodological development would be useful

Conclusions

- We have identified four sources of systematics in unfolding:
  1. Regularization bias
  2. Wide-bin bias
  3. Missing nuisance variables
  4. Uncertainty in the response kernel k
- Handling the regularization bias has been a long-time challenge in unfolding
- 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. (2021) for methods and simulation results
- Handling known nuisance variables is "easy", but what to do about unknown nuisance variables?
- Interesting open methodological questions about handling the uncertainty in the response kernel k

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

C. Genovese and L. Wasserman, Adaptive confidence bands, *The Annals of Statistics*, 36 (2):875–905, 2008.

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.

References III

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

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.

B. W. Rust and W. R. Burrus. *Mathematical Programming and the Numerical Solution of Linear Equations*. American Elsevier, 1972.

B. W. Rust and D. P. O'Leary, Confidence intervals for discrete approximations to ill-posed problems, *Journal of Computational and Graphical Statistics*, 3(1):67–96, 1994.

References IV

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, arXiv:2111.01091 [stat.AP], 2021.

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

Relaxing the full rank assumption

The simple least-squares approach works as long as the forward model $\boldsymbol{K}$ has full column rank and there are no constraints that $\boldsymbol{x}$ needs to satisfy

The full rank requirement can be quite restrictive in practice, for example:

- Unfolding with more true bins p than smeared bins $n \Rightarrow \boldsymbol{K}$ column-rank deficient

When $\boldsymbol{K}$ is column-rank deficient, it has a non-trivial null space $\ker(\boldsymbol{K})$

Therefore, confidence intervals for $\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$ would need to be infinitely long if there are no constraints on $\boldsymbol{x}$ (assuming $\boldsymbol{h}$ not orthogonal to $\ker(\boldsymbol{K})$)

However, simple constraints such as $\boldsymbol{x} \geq \boldsymbol{0}$ or $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ can be enough to make the intervals finite

- And we would in any case like to make use of constraints, if available

Strict bounds: Motivation

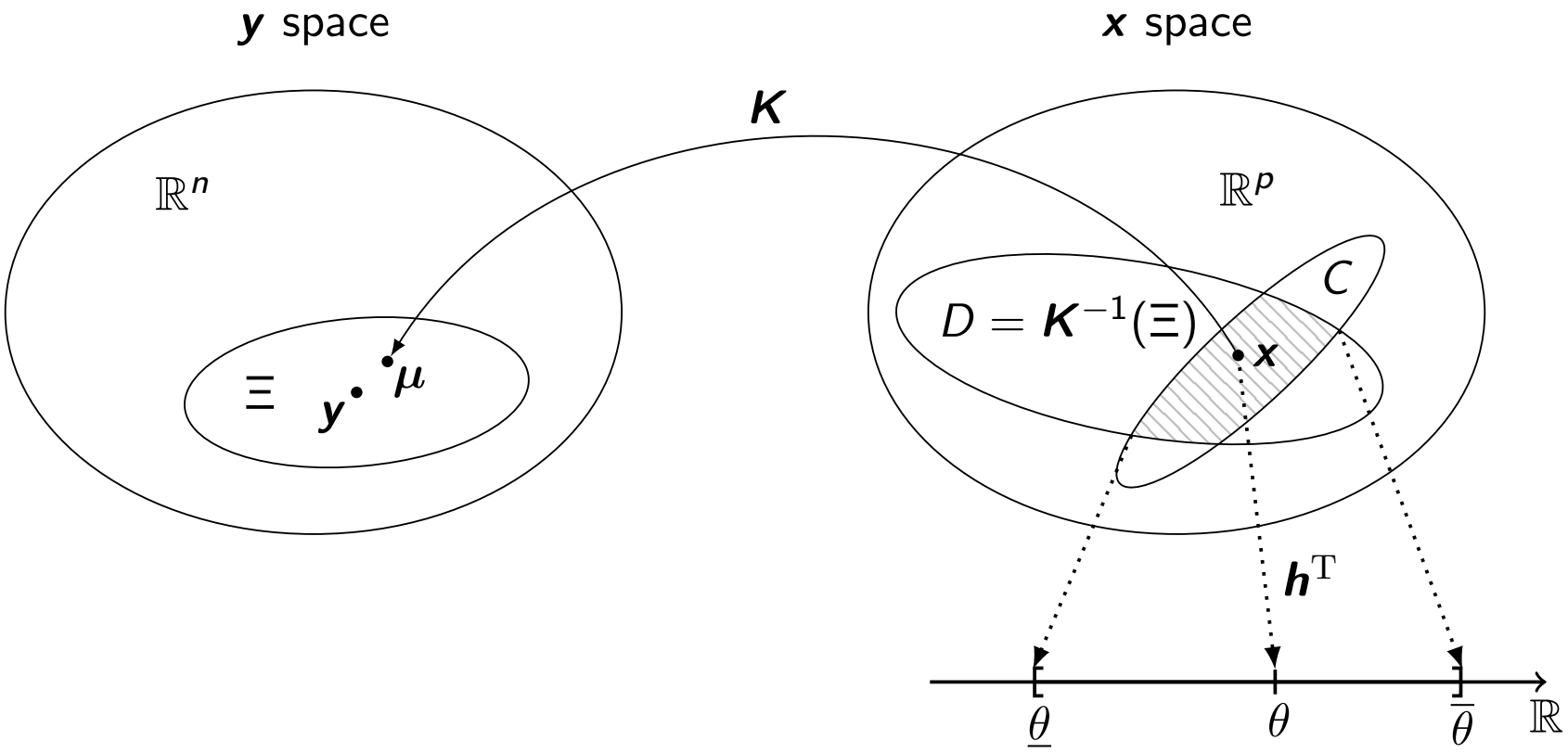
So the question becomes:

> Assuming model $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{K}$ need not have full column rank, how does one obtain a finite-sample $1 - \alpha$ confidence interval for the linear functional $\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$ subject to the constraint $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$?

In the following:

- We assume that we have transformed the problem so that $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I})$
- We denote the noise-free data by $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{x}$

Strict bounds (e.g., Stark (1992))



$$\theta = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}, \quad \underline{\theta} = \min_{\boldsymbol{x} \in C \cap D} \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}, \quad \overline{\theta} = \max_{\boldsymbol{x} \in C \cap D} \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$$

$$\begin{aligned} \mathrm{P}(\boldsymbol{\mu} \in \Xi) \geq 1 - \alpha \quad &\Rightarrow \quad \mathrm{P}(\boldsymbol{x} \in D) \geq 1 - \alpha \\ &\Rightarrow \quad \mathrm{P}(\boldsymbol{x} \in C \cap D) \geq 1 - \alpha \\ &\Rightarrow \quad \mathrm{P}(\theta \in [\underline{\theta}, \overline{\theta}]) \geq 1 - \alpha \end{aligned}$$

Strict bounds

If we construct the confidence set Ξ as

$$\Xi = \{\boldsymbol{\mu} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{\mu}\|^2 \leq \chi^2_{n,1-\alpha}\},$$

then the endpoints of the confidence interval for θ are given by the solutions of the following two quadratic programs:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq \chi^2_{n,1-\alpha} \\ & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{array}$$

and

$$\begin{array}{ll} \text{maximize} & \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq \chi^2_{n,1-\alpha} \\ & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{array}$$

The resulting interval $[\underline{\theta}, \overline{\theta}]$ has *by construction* coverage at least $1 - \alpha$

Strict bounds

The main limitation of the previous construction is that there is slack in the last step:

$$\mathrm{P}(\boldsymbol{x} \in C \cap D) \geq 1 - \alpha \quad \Rightarrow \quad \mathrm{P}(\theta \in [\underline{\theta}, \overline{\theta}]) \geq 1 - \alpha$$

Because $C \cap D$ is a simultaneous confidence set for $\boldsymbol{x}$, this construction yields *simultaneous* confidence intervals for any arbitrarily large collection of functionals of $\boldsymbol{x}$

This means that, if evaluated as a *one-at-a-time* interval, $[\underline{\theta}, \overline{\theta}]$ from this construction is necessarily conservative (i.e., it has overcoverage)

One-at-a-time strict bounds

It has been conjectured (Rust and Burrus, 1972; Rust and O'Leary, 1994) that the following modification gives a shorter but still valid one-at-a-time interval:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + s^2 \\ & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{array}$$

and

$$\begin{array}{ll} \text{maximize} & \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + s^2 \\ & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{array}$$

where $s^2 = \min_{\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2$

Decision-theoretic UQ for inverse problems

The one-at-a-time strict bounds have excellent empirical coverage but, as of now, we do not have a proof of coverage and/or optimality

However, there is an alternative approach we can take, where we can prove coverage and where the intervals are in a certain sense optimal

As a by-product of the previous analysis, we showed that the following interval has coverage at least $1-\alpha$:

$$\left[\underline{\theta}, \overline{\theta}\right] = \left[\underline{\boldsymbol{w}}^T \boldsymbol{y} - z_{1-\alpha/2}\|\underline{\boldsymbol{w}}\| - \boldsymbol{b}^T \underline{\boldsymbol{c}},\ \overline{\boldsymbol{w}}^T \boldsymbol{y} + z_{1-\alpha/2}\|\overline{\boldsymbol{w}}\| + \boldsymbol{b}^T \overline{\boldsymbol{c}}\right],$$

when $(\underline{\boldsymbol{w}}, \underline{\boldsymbol{c}})$ satisfies $\boldsymbol{h} + \boldsymbol{A}^T \underline{\boldsymbol{c}} - \boldsymbol{K}^T \underline{\boldsymbol{w}} = \boldsymbol{0}$, $\underline{\boldsymbol{c}} \geq \boldsymbol{0}$ and $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{c}})$ satisfies $\boldsymbol{h} - \boldsymbol{A}^T \overline{\boldsymbol{c}} - \boldsymbol{K}^T \overline{\boldsymbol{w}} = \boldsymbol{0}$, $\overline{\boldsymbol{c}} \geq \boldsymbol{0}$

The coverage is guaranteed for *any* $(\underline{\boldsymbol{w}}, \underline{\boldsymbol{c}})$ and $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{c}})$ that satisfy the above constraints and do not depend on $\boldsymbol{y}$

Decision-theoretic UQ for inverse problems

So we can take a decision-theoretic approach and optimize the expected length of the interval $\left[\underline{\theta}, \overline{\theta}\right]$ with respect to a prior distribution on $\boldsymbol{x}$

Consider the lower bound

$$\underline{\theta} = \underline{\boldsymbol{w}}^T \boldsymbol{y} - z_{1-\alpha/2} \|\underline{\boldsymbol{w}}\| - \boldsymbol{b}^T \underline{\boldsymbol{c}}$$

The expectation of $\underline{\theta}$ under the noise ε is

$$\mathrm{E}_{\varepsilon}(\underline{\theta}) = \underline{\boldsymbol{w}}^T \boldsymbol{K}\boldsymbol{x} - z_{1-\alpha/2} \|\underline{\boldsymbol{w}}\| - \boldsymbol{b}^T \underline{\boldsymbol{c}} = R(\boldsymbol{x})$$

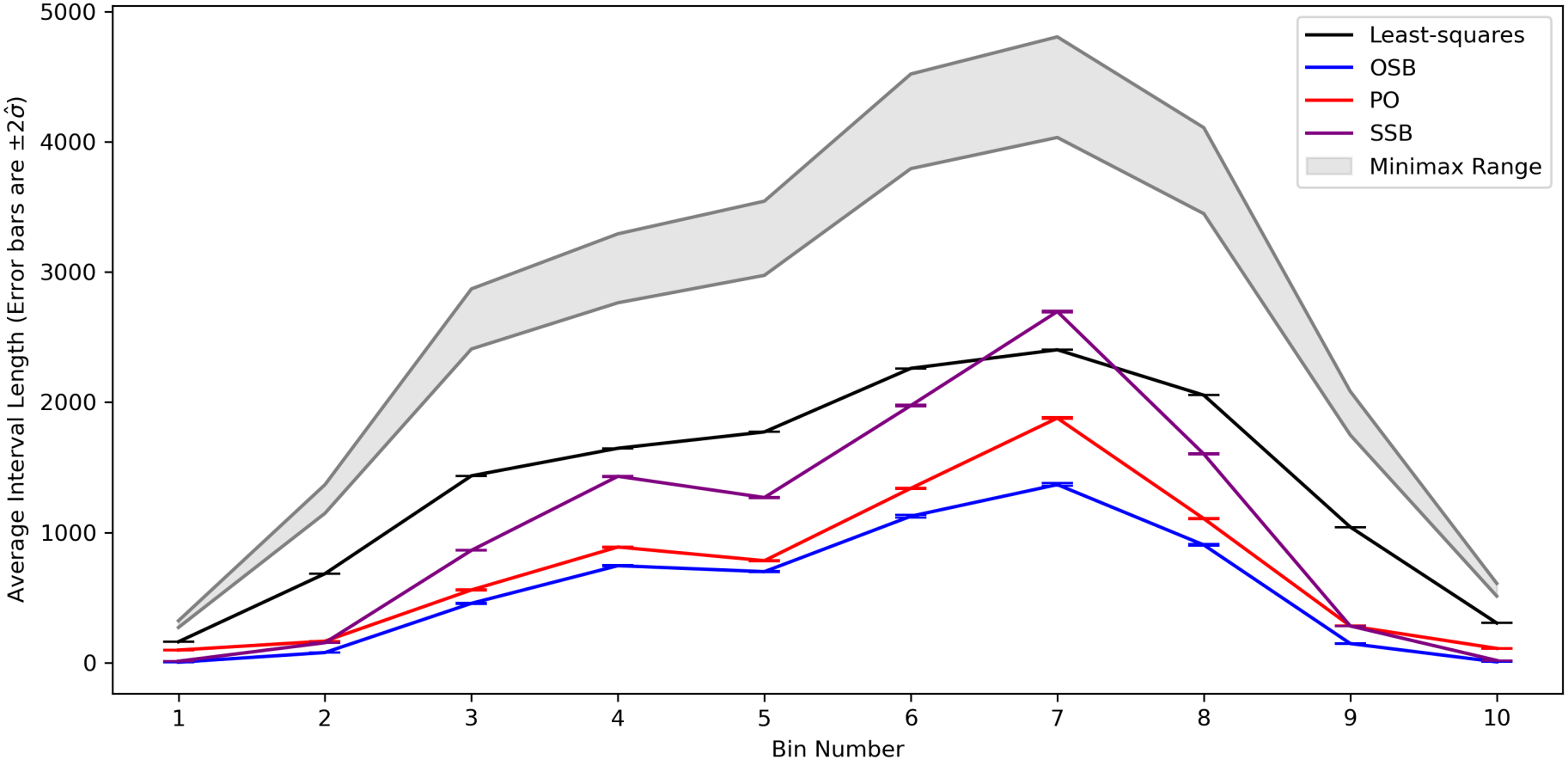
If we assume a prior on $\boldsymbol{x}$ with expectation $\boldsymbol{x}_a$, we can further compute the expectation of $R(\boldsymbol{x})$ w.r.t. $\boldsymbol{x}$:

$$\mathrm{E}_{\boldsymbol{x}}(R(\boldsymbol{x})) = \underline{\boldsymbol{w}}^T \boldsymbol{K}\boldsymbol{x}_a - z_{1-\alpha/2} \|\underline{\boldsymbol{w}}\| - \boldsymbol{b}^T \underline{\boldsymbol{c}}$$

Maximizing the above subject to the constraints on $(\underline{\boldsymbol{w}}, \underline{\boldsymbol{c}})$ will optimize the lower bound $\underline{\theta}$; we can then do an analogous calculation to optimize the upper bound $\overline{\theta}$

The result is an interval $\left[\underline{\theta}, \overline{\theta}\right]$ with coverage at least $1 - \alpha$ whose expected length $\mathrm{E}_{\boldsymbol{x},\varepsilon}(\overline{\theta} - \underline{\theta})$ is minimized for the assumed prior on $\boldsymbol{x}$

Prior-optimized confidence intervals for wide-bin unfolding



(Joint work with Michael Stanley and Pratik Patil)

Properties of the different intervals

Interval Type	Coverage Design	Interval Width	Empirical Coverage	Provable Coverage	Physical Constraints	Rank-Deficient Model
Tikhonov/Regularized	One-at-a-time	Narrow	✗	✗	✗	✓
Least-Squares	One-at-a-time	Medium	✓*	✓	✗	✗
OSB	One-at-a-time	Medium	✓	?	✓	✓
PO	One-at-a-time	Medium	✓	✓	✓**	✓
SSB	Simultaneous	Wide	✓	✓	✓	✓
Minimax	One-at-a-time	Wide	✓*	✓	✓	✗

Regularized unfolding

Two popular approaches to regularized unfolding:

1. Tikhonov regularization (Höcker and Kartvelishvili, 1996; Schmitt, 2012)
2. Expectation-maximization iteration with early stopping (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`[3] (Schmitt, 2012): $P(\boldsymbol{\lambda}) = \|\mathsf{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\mathrm{MC}})\|^2$

[3] `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, $k = 0$

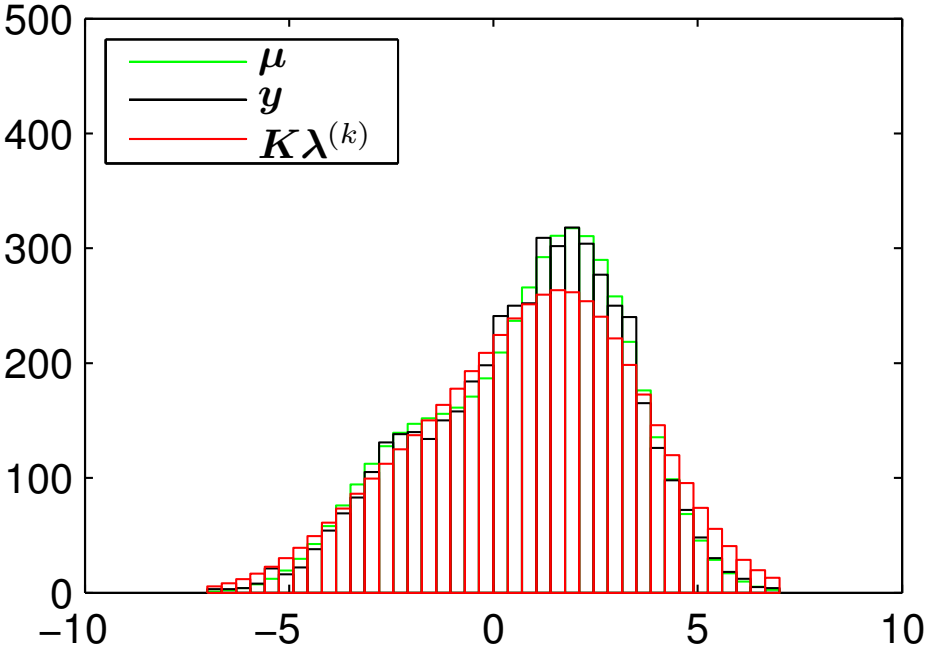


Figure: Smeared histogram

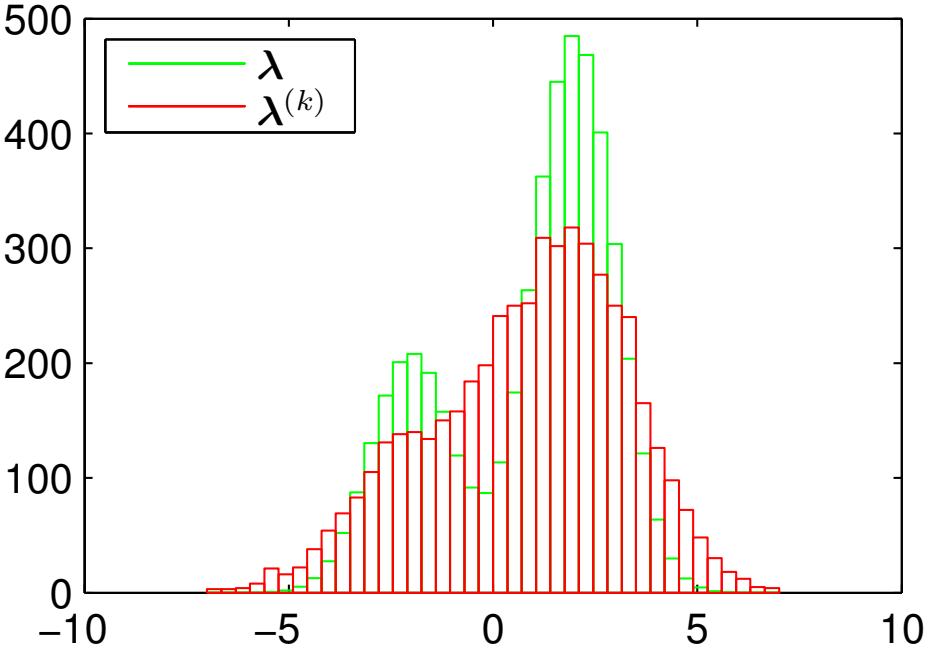


Figure: True histogram

D'Agostini demo, $k = 100$

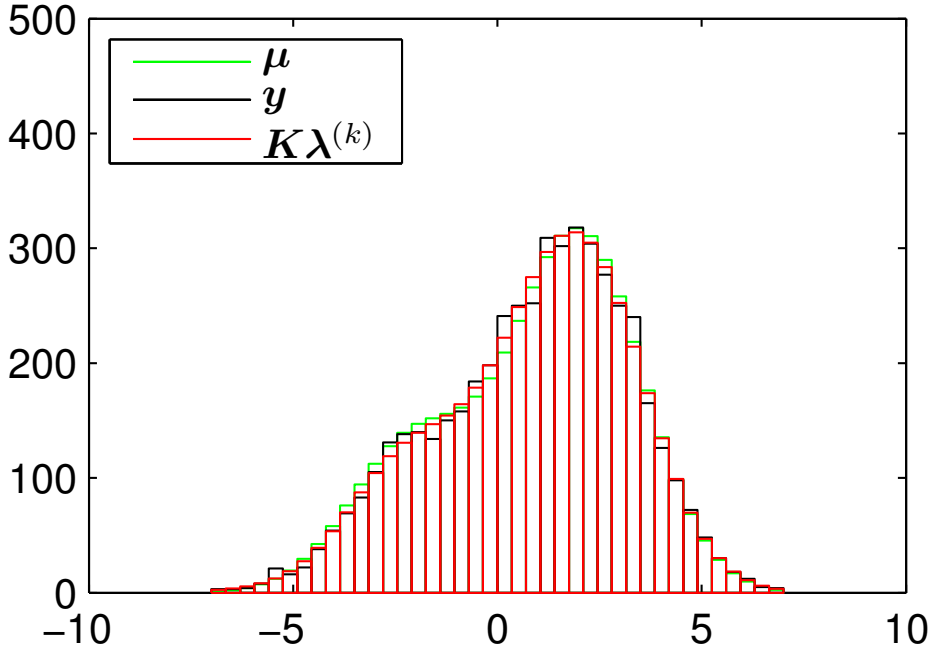


Figure: Smeared histogram

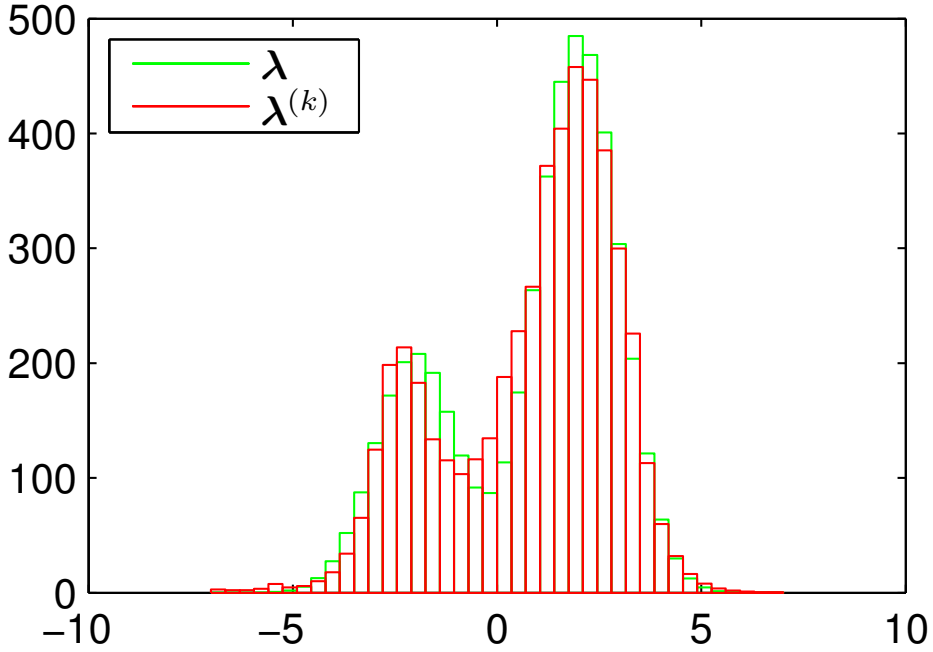


Figure: True histogram

D'Agostini demo, $k = 10000$

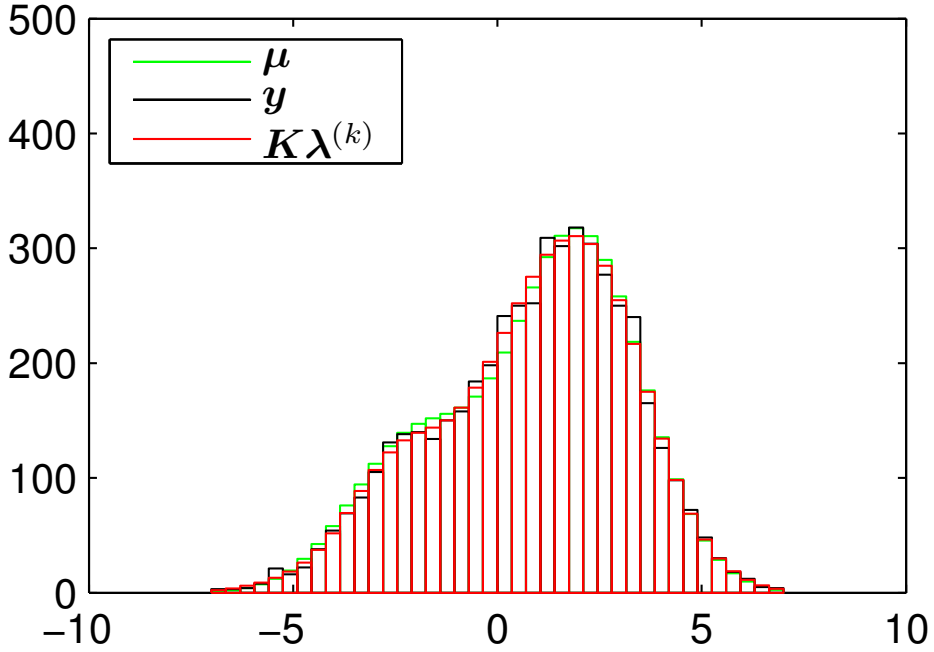


Figure: Smeared histogram

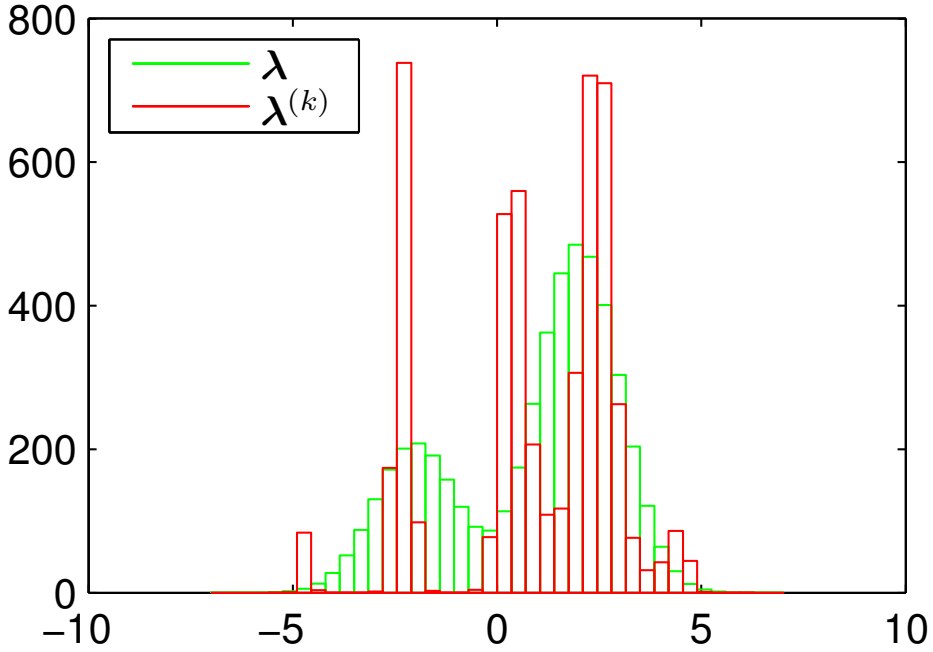


Figure: True histogram

D'Agostini demo, $k = 100000$

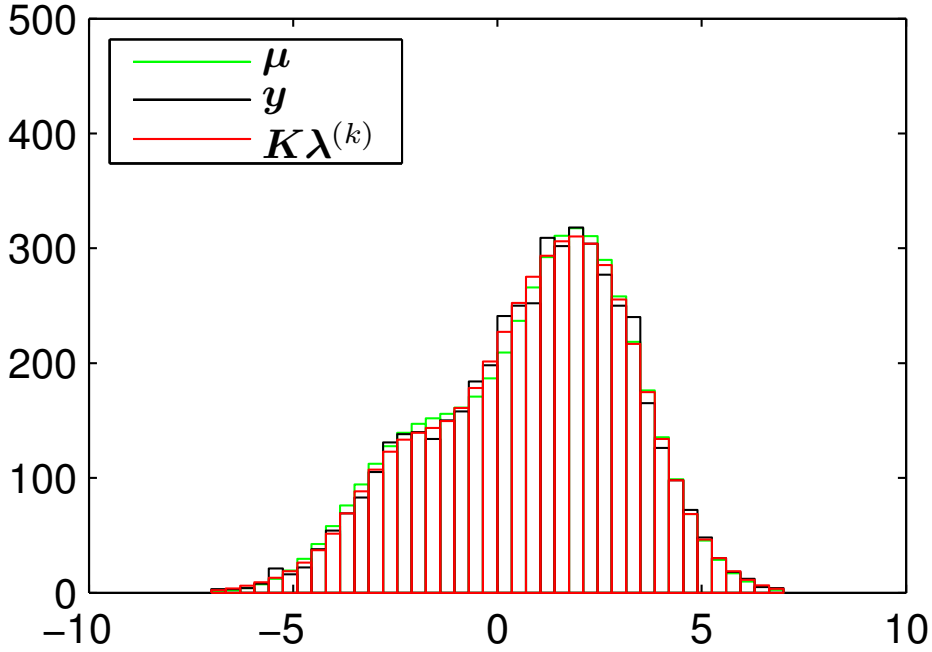


Figure: Smeared histogram

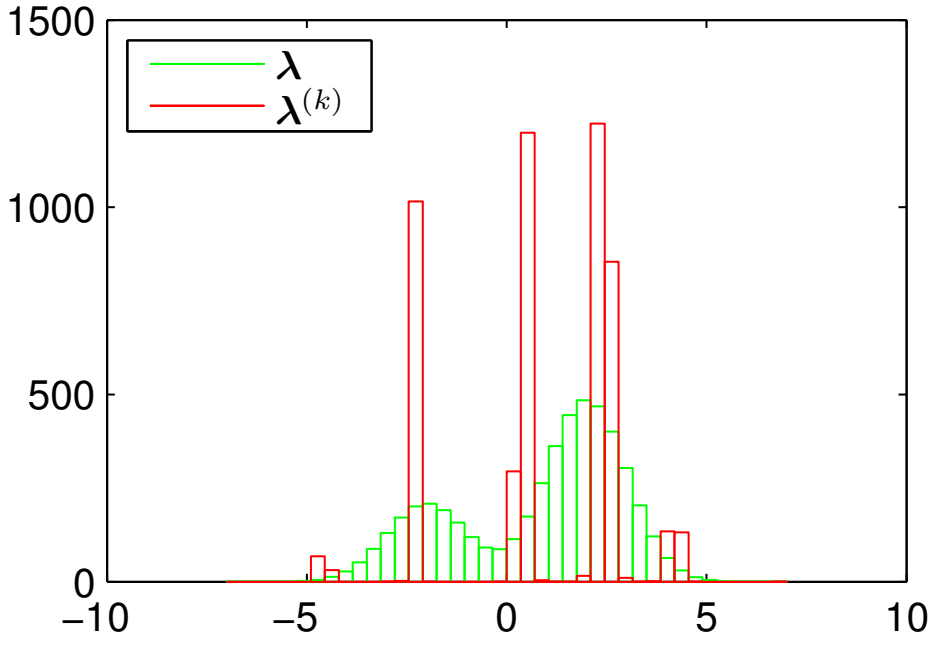


Figure: True histogram

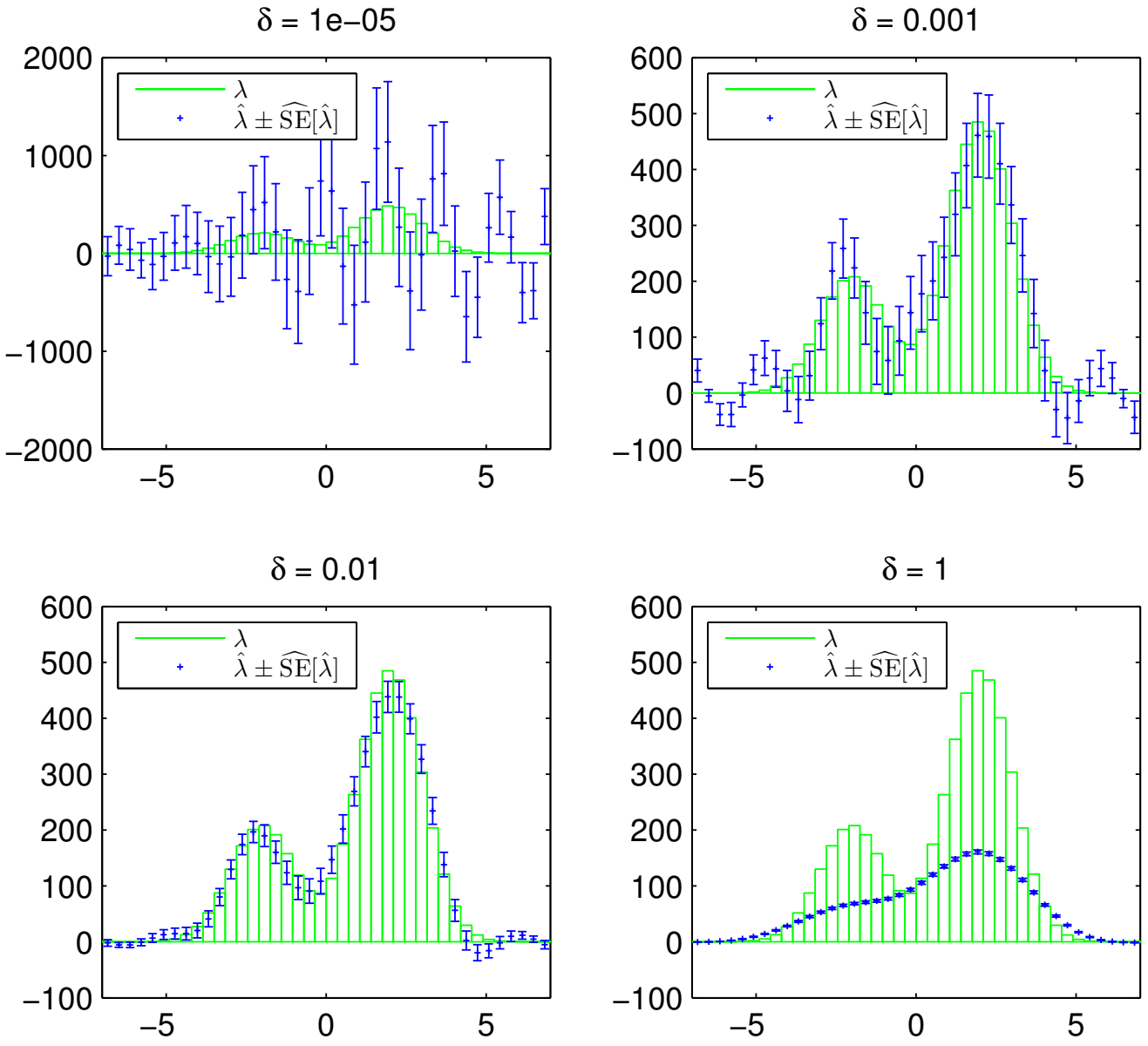
Other methods

- **Bin-by-bin correction factors**
  - Attempts to unfold resolution effects by performing multiplicative efficiency corrections
  - This method is simply wrong and must not be used
- **Fully Bayesian unfolding (FBU)** (Choudalakis, 2012)
  - Unfolding using Bayesian statistics where the prior regularizes the ill-posed problem
  - Certain priors lead to solutions similar to Tikhonov, but with Bayesian credible intervals as the uncertainties
  - Note: D'Agostini has nothing to do with proper Bayesian inference
- **Gaussian processes** (Bozson et al., 2018; Stuart, 2010)
  - Very closely related to Tikhonov regularization / penalized maximum likelihood / FBU
  - Inherits many of the same limitations
- **RUN/TRUEE** (Blobel, 1985, 1996; Milke et al., 2013)
  - Penalized maximum likelihood with B-spline discretization
- **Shape-constrained unfolding** (Kuusela and Stark, 2017)
  - Correct-coverage simultaneous confidence intervals by imposing constraints on positivity, monotonicity and/or convexity
- **Expectation-maximization with smoothing** (Volobouev, 2015)
  - Adds a smoothing step to each iteration of D'Agostini and iterates until convergence
- **Iterative dynamically stabilized unfolding** (Malaescu, 2011)
  - "The regularisation method in IDS is local, based on the statistical significance of the data-MC differences in each bin." (B. Malaescu)
  - I have not seen this used in CMS, but it seems to be quite popular in ATLAS
- ...

Choice of the regularization strength

- A key issue in unfolding is the choice of the regularization strength (δ in Tikhonov, # of iterations in D'Agostini)
  - The solution and especially the uncertainties depend heavily on this choice
- This choice should 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
- Note: *All of these are designed for point estimation!*
  - Not necessarily optimal for uncertainty quantification

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



Uncertainty quantification in unfolding

- Aim: Find random intervals $\left[\underline{\lambda}_i(\boldsymbol{y}), \overline{\lambda}_i(\boldsymbol{y})\right]$, $i = 1, \ldots, p$, with *coverage probability* $1 - \alpha$:

$$\mathsf{P}\big(\lambda_i \in \left[\underline{\lambda}_i(\boldsymbol{y}), \overline{\lambda}_i(\boldsymbol{y})\right]\big) \approx 1 - \alpha$$

- Most implementations quantify the uncertainty using the binwise variance (estimated using either error propagation or resampling):

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

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

Undersmoothed unfolding

- Standard methods for picking the regularization strength optimize the point estimation performance
  - These estimators have 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*
- I have been working with Lyle Kim to implement 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, $\boldsymbol{\lambda}^{\mathrm{MC}} = 0$

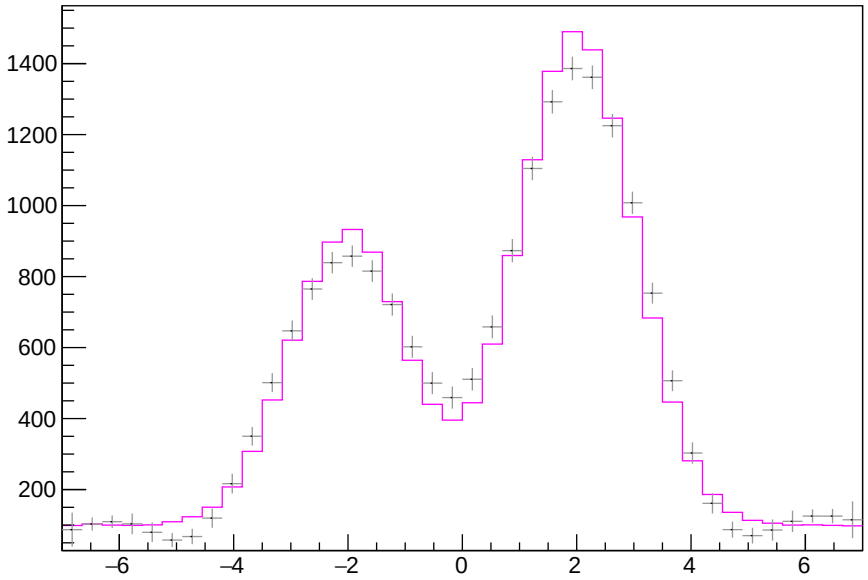


Figure: L-curve, $\tau = \sqrt{\delta}$=0.01186

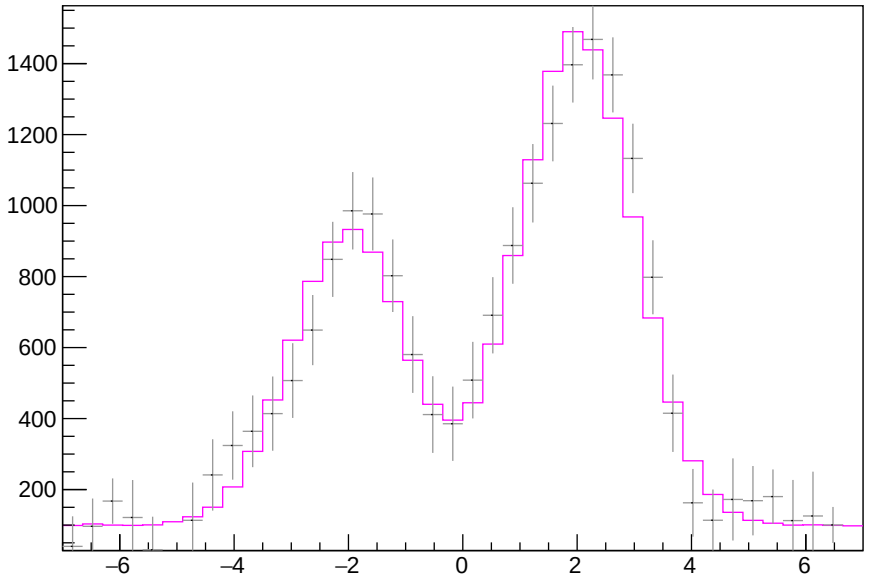


Figure: Undersmoothing, $\tau = \sqrt{\delta}$=0.00177

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

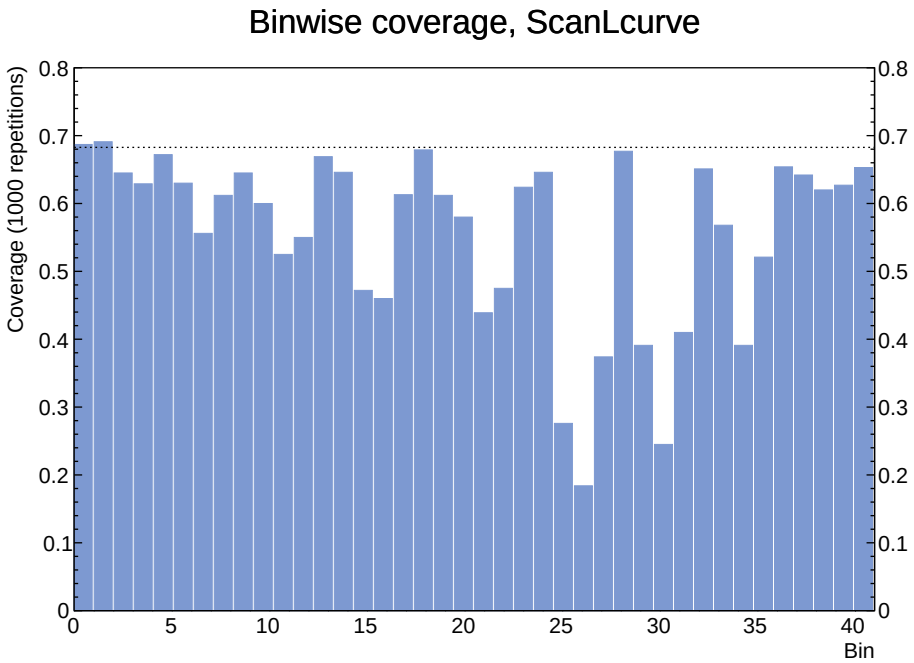


Figure: L-curve

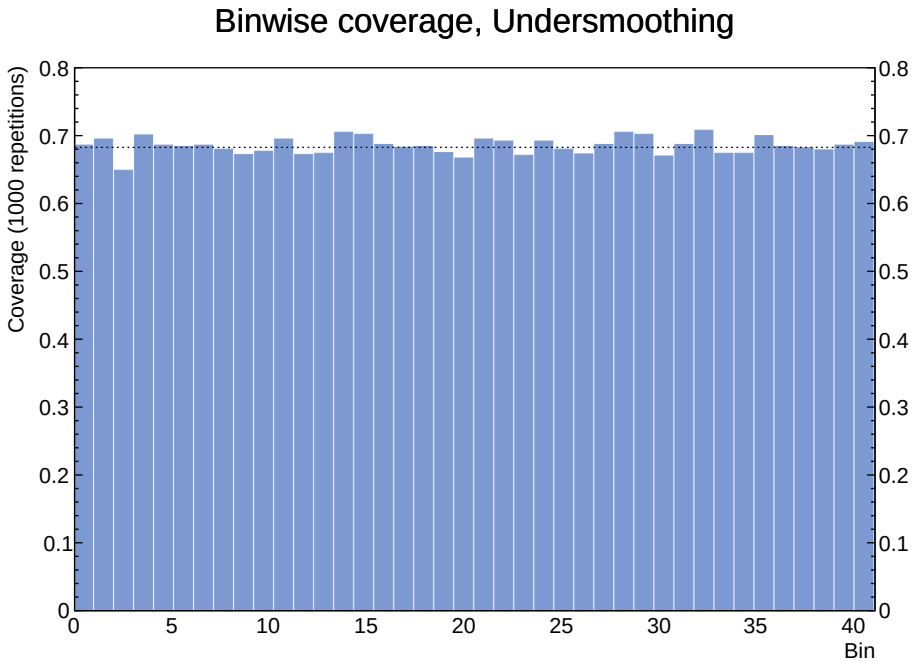


Figure: Undersmoothing

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

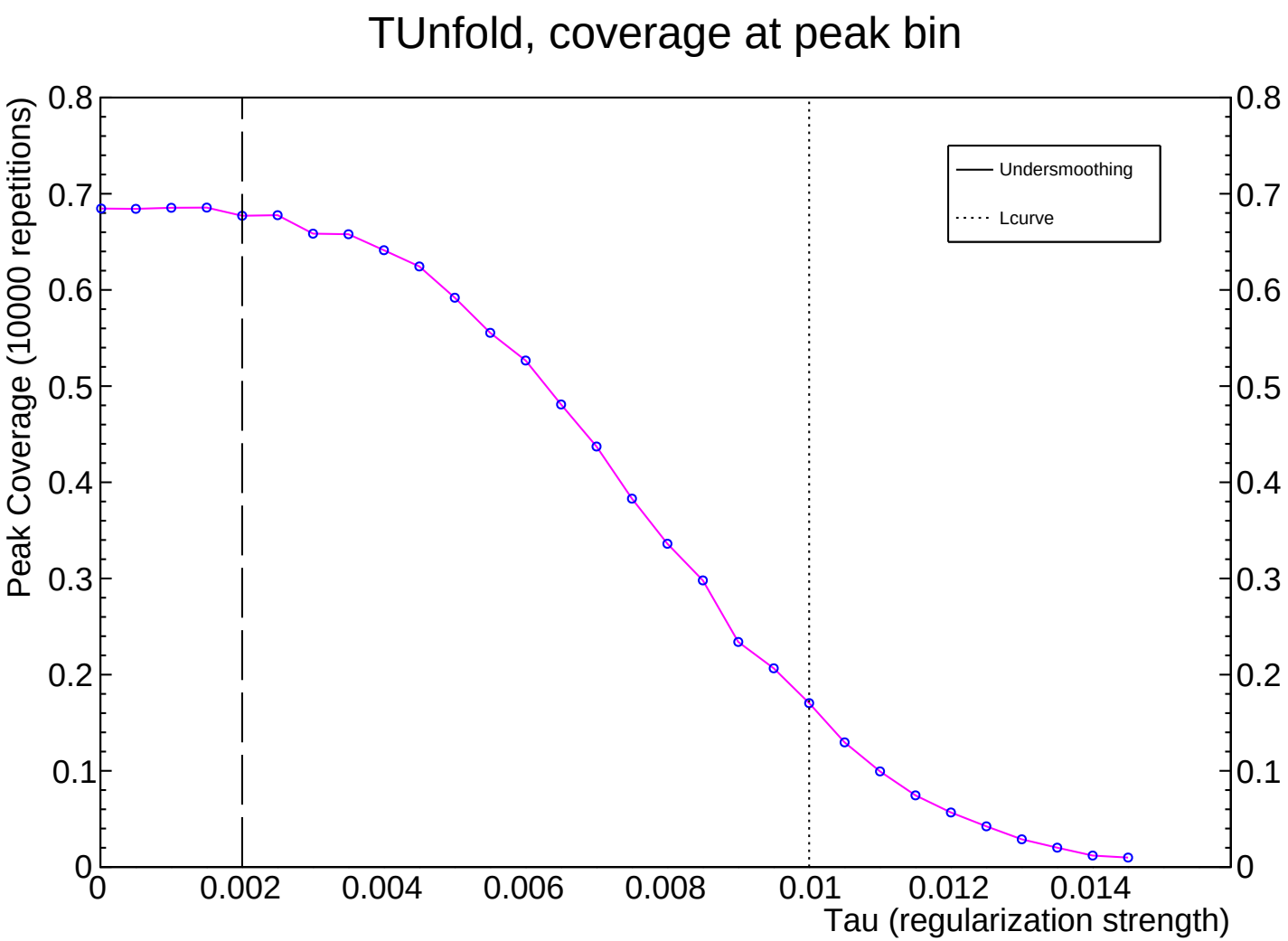
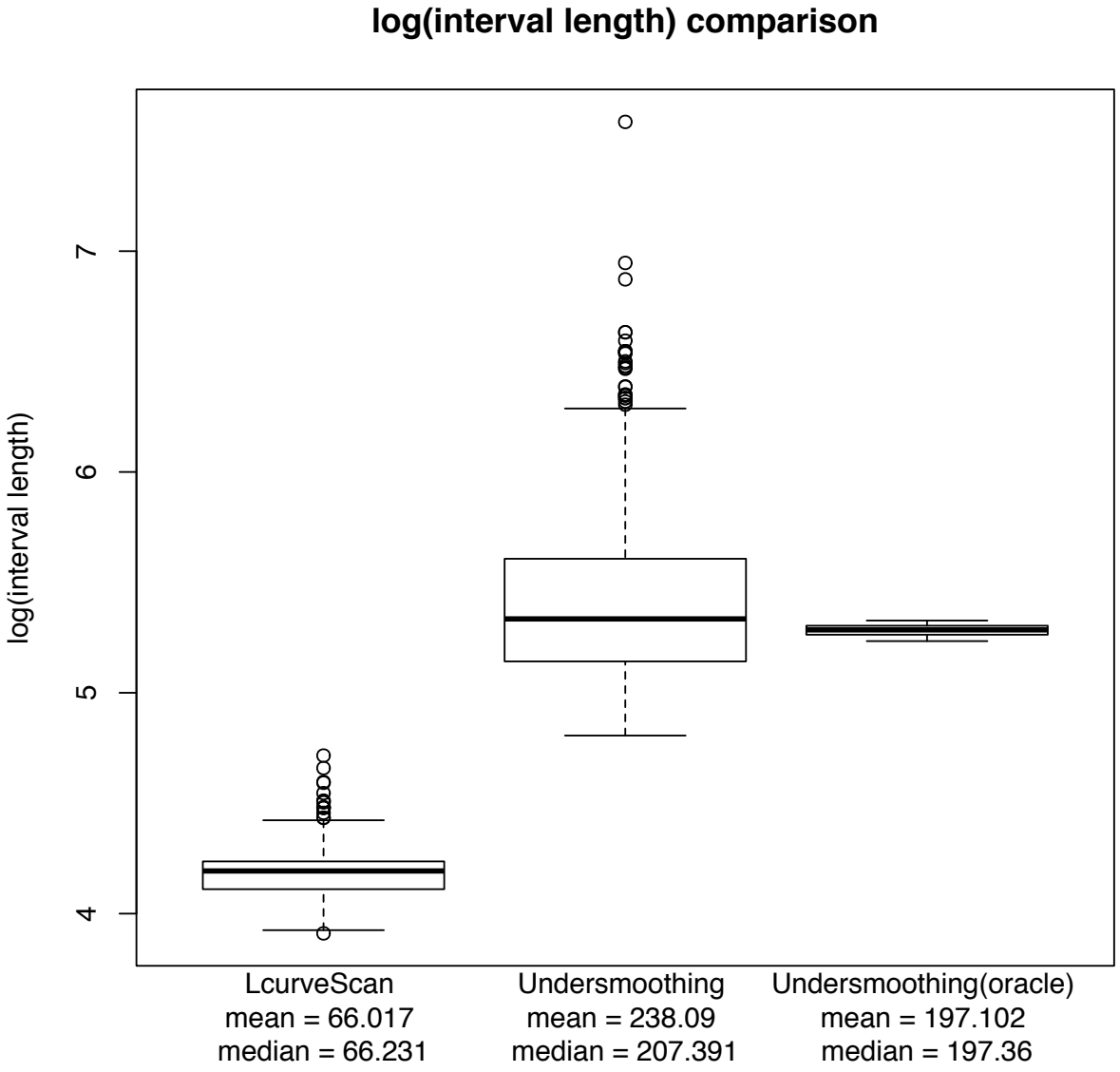
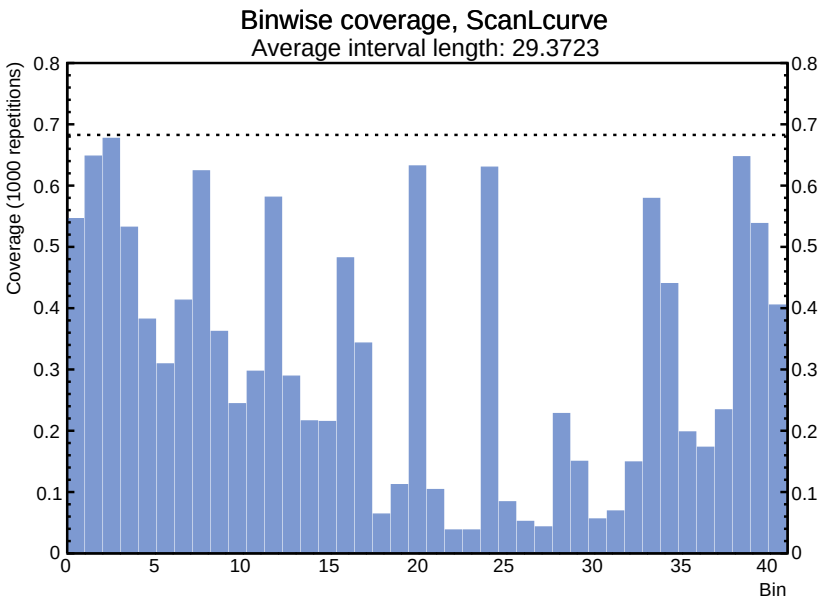


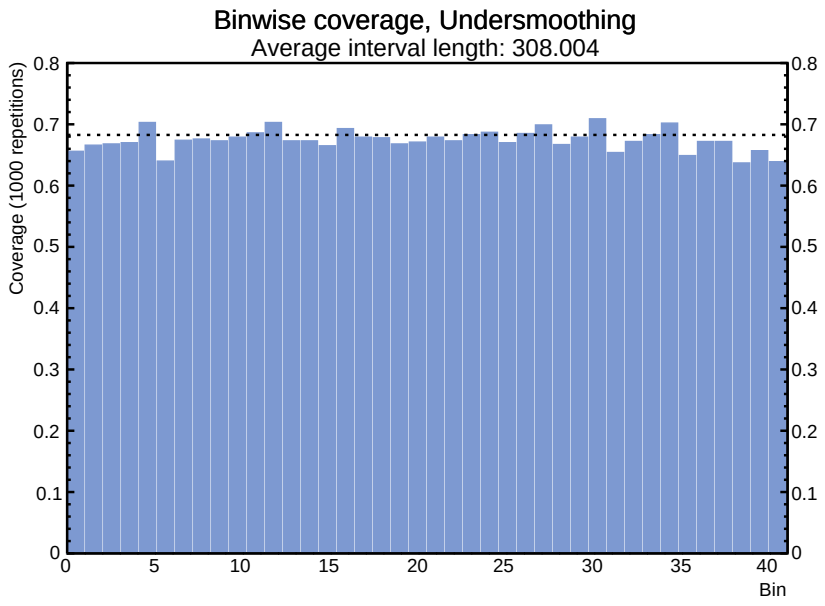
Figure: Coverage at the right peak of a bimodal density

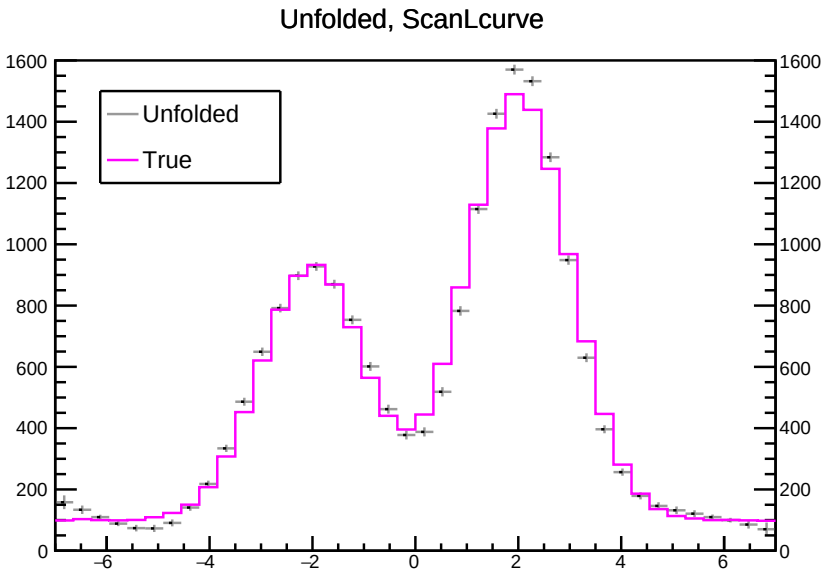
Interval lengths, $\boldsymbol{\lambda}^{\mathrm{MC}} = 0$

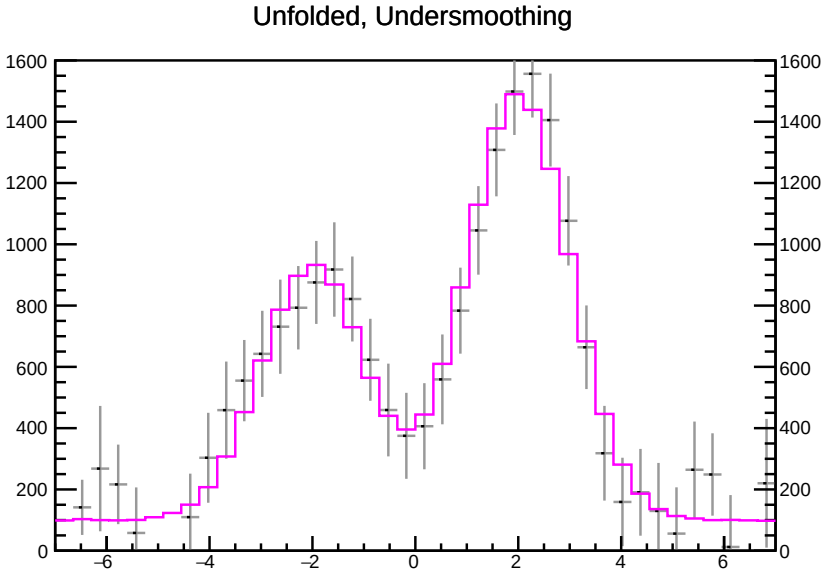


Histograms, coverage and interval lengths when $\boldsymbol{\lambda}^{\mathrm{MC}} \neq 0$









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