Introduction to Unfolding Methods in High Energy Physics

Mikael Kuusela

Department of Statistics and Data Science,
Carnegie Mellon University

Pan-European Advanced School on Statistics in High Energy Physics

DESY, Hamburg, Germany

October 31, 2019

Outline

Outline

The unfolding problem

- Unfolding refers to the problem of estimating the particle-level spectrum of some physical quantity of interest on the basis of observations smeared by an imperfect measurement device
- What would the distribution look like when measured with a device having a perfect experimental resolution?
    - Cf. deconvolution in optics, image reconstruction in medical imaging

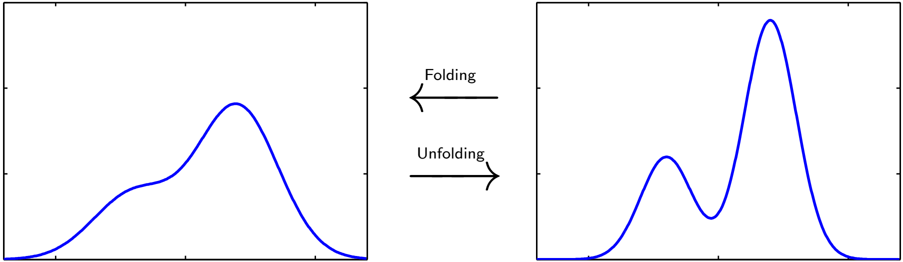


Figure: Smeared spectrum

Figure: True spectrum

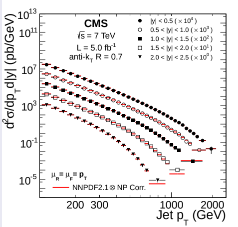
Why unfold?

Unfolding is usually done to achieve one or more of the following goals:

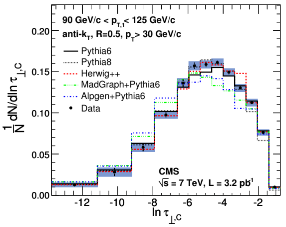
1. **Comparison of experiments with different responses**
2. **Comparison of the measurement with future theories**
   $\rightarrow$ Controversial since you could also think of smearing the theory
3. **Input to a subsequent analysis**
4. **Exploratory data analysis**

Examples of unfolding in LHC data analysis

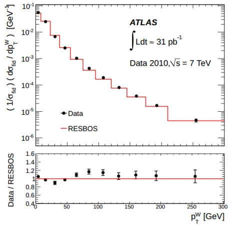
Inclusive jet cross section



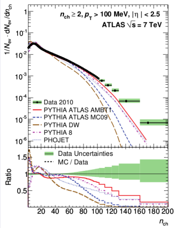
Hadronic event shape



W boson cross section



Charged particle multiplicity



Problem formulation

- Notation:
  - $\boldsymbol{\lambda} \in \mathbb{R}_+^p$ bin means of the true histogram
  - $\mathbf{x} \in \mathbb{N}_0^p$ bin counts of the true histogram
  - $\boldsymbol{\mu} \in \mathbb{R}_+^n$ bin means of the smeared histogram
  - $\mathbf{y} \in \mathbb{N}_0^n$ bin counts of the smeared histogram
- Assume that:
  1. The true counts are independent and Poisson distributed

$$\mathbf{x}|\boldsymbol{\lambda} \sim \mathrm{Poisson}(\boldsymbol{\lambda}), \quad \perp\!\!\!\perp\, x_i|\boldsymbol{\lambda}$$

  2. The propagation of events to neighboring bins is multinomial conditional on x_i and independent for each true bin
- It follows that the smeared counts are also independent and Poisson distributed

$$\mathbf{y}|\boldsymbol{\lambda} \sim \mathrm{Poisson}(\mathbf{K}\boldsymbol{\lambda}), \quad \perp\!\!\!\perp\, y_i|\boldsymbol{\lambda}$$

Problem formulation

- Here the elements of the *response matrix* $\mathbf{K} \in \mathbb{R}^{n \times p}$ are given by

$$K_{ij} = P(\text{smeared event in bin } i \,|\, \text{true event in bin } j)$$

  and assumed to be known
- $\boldsymbol{K}$ relates the smeared mean $\boldsymbol{\mu}$ and the true mean $\boldsymbol{\lambda}$ as $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$
- The unfolding problem:

Problem statement

Given the smeared observations $\mathbf{y}$ and the Poisson regression model

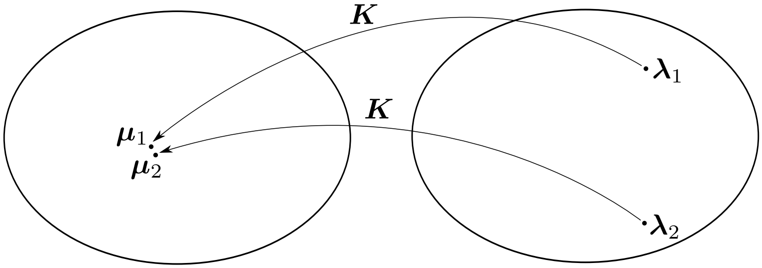
$$\mathbf{y}|\boldsymbol{\lambda} \sim \mathrm{Poisson}(\mathbf{K}\boldsymbol{\lambda}),$$

what can be said about the means $\boldsymbol{\lambda}$ of the true histogram?

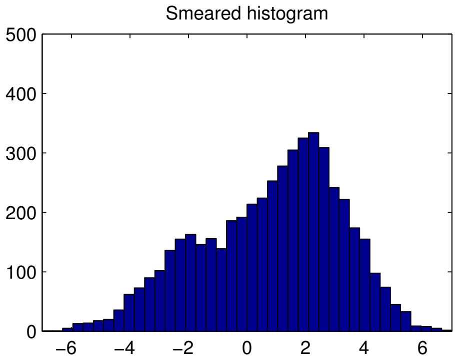
- The problem here is that typically $\mathbf{K}$ is an ill-conditioned matrix

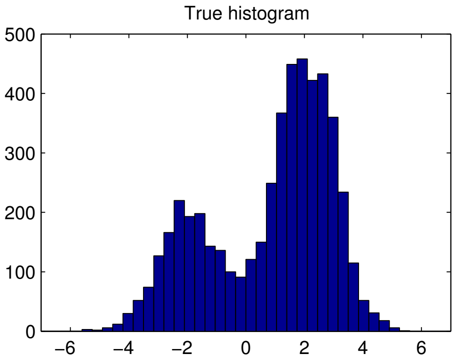
Unfolding is an ill-posed inverse problem

- The linear system $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ is typically ill-conditioned
  - That is, true histograms $\boldsymbol{\lambda}$ that are very different can map into smeared histograms $\boldsymbol{\mu}$ that are very similar
- As a result, the (pseudo)inverse of $\boldsymbol{K}$ is very sensitive to statistical fluctuations in the smeared data



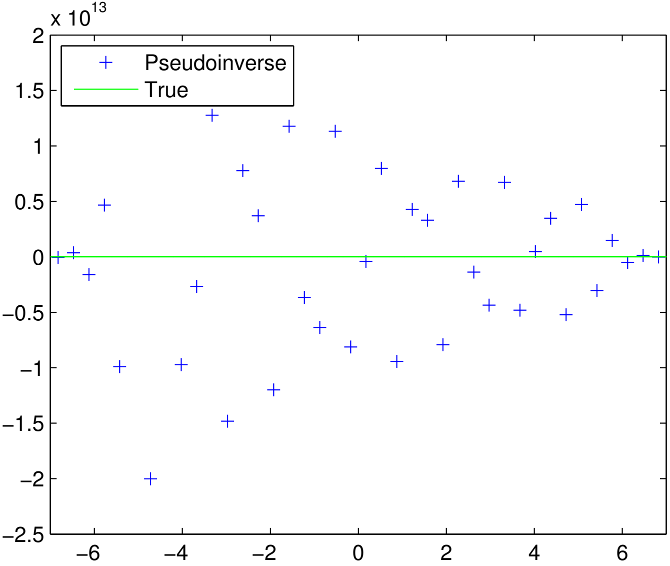
Demonstration of ill-posedness





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

Demonstration of ill-posedness



Outline

1. Introduction

2. Basic unfolding methodology
   - Maximum likelihood estimation
   - Regularized frequentist techniques
   - Bayesian unfolding

3. Challenges and opportunities in unfolding
   - Choice of the regularization strength
   - Uncertainty quantification
   - MC dependence of the response matrix
   - Wide-bin unfolding

4. Conclusions

Outline

The likelihood function

- The *likelihood function* in unfolding is:

$$L(\boldsymbol{\lambda}) = p(\mathbf{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\lambda}) = \prod_{i=1}^{n} \frac{\left(\sum_{j=1}^{p} K_{ij}\lambda_j\right)^{y_i}}{y_i!} e^{-\sum_{j=1}^{p} K_{ij}\lambda_j}, \quad \boldsymbol{\lambda} \in \mathbb{R}_+^p$$

- This function uses our Poisson regression model to link the observations $\mathbf{y}$ with the unknown $\boldsymbol{\lambda}$
  - The likelihood function plays a key role in all sensible unfolding methods
- In most statistical problems, the maximum of the likelihood (or equivalently the maximum of the log-likelihood) provides a good estimate of the unknown
  - In ill-posed problems, *this is usually not the case*, but the maximum likelihood solution still provides a good starting point

Maximum likelihood estimation

- Any histogram that maximizes the log-likelihood of the unfolding problem is called a *maximum likelihood estimator* $\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}$ of $\boldsymbol{\lambda}$
- Hence, we want to solve:

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} \log p(\mathbf{y}|\boldsymbol{\lambda}) = \sum_{i=1}^{n} \left[y_i \log \left(\sum_{j=1}^{p} K_{ij} \lambda_j \right) - \sum_{j=1}^{p} K_{ij} \lambda_j \right] + \mathrm{const}$$

- How to find the maximizer?

Maximum likelihood estimation

Proposition

Let $\mathbf{K}$ be an invertible square matrix and assume that $\hat{\boldsymbol{\lambda}} = \mathbf{K}^{-1}\mathbf{y} \geq \mathbf{0}$. Then $\hat{\boldsymbol{\lambda}}$ is the MLE of $\boldsymbol{\lambda}$.

- That is, matrix inversion gives us the MLE if $\mathbf{K}$ is invertible and the resulting estimate is positive
- Note that this result is more restrictive than it may seem
  - $\mathbf{K}$ is often non-square
  - Even if $\mathbf{K}$ was square, it is often not invertible
  - And even if $\mathbf{K}$ was invertible, $\mathbf{K}^{-1}\mathbf{y}$ often contains negative values
- Is there a general recipe for finding the MLE?

Maximum likelihood estimation

- The MLE can always be found computationally by using the *expectation-maximization (EM) algorithm* (Dempster et al. (1977))
  - This is a widely used iterative algorithm for finding maximum likelihood solutions in problems that can be seen as containing incomplete observations
- Starting from some initial value $\boldsymbol{\lambda}^{(0)} > \mathbf{0}$, the EM iteration for unfolding is given by:

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

- The convergence of this iteration to an MLE (i.e. $\boldsymbol{\lambda}^{(k)} \overset{k\to\infty}{\longrightarrow} \hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}$) was proved by Vardi et al. (1985)

Maximum likelihood estimation

- The EM iteration for finding the MLE in Poisson regression problems has been rediscovered many times in different fields:
  - Optics: Richardson (1972)
  - Astronomy: Lucy (1974)
  - Tomography: Shepp and Vardi (1982); Lange and Carson (1984); Vardi et al. (1985)
  - HEP: Kondor (1983); Mülthei and Schorr (1987b,a, 1989); D'Agostini (1995)
- In modern use, the algorithm is most often called *D'Agostini iteration* in particle physics and *Lucy–Richardson deconvolution* in astronomy and optics
- In particle physics, also the name "Bayesian unfolding" has been used but this is an unfortunate misnomer
  - D'Agostini iteration is a fully frequentist technique for finding the MLE
  - *There is nothing Bayesian about it!*

D'Agostini demo, $k = 0$

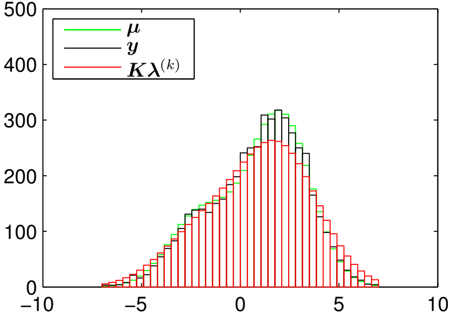


Figure: Smeared histogram

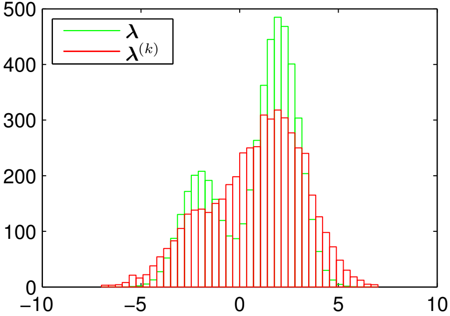


Figure: True histogram

D'Agostini demo, $k = 100$

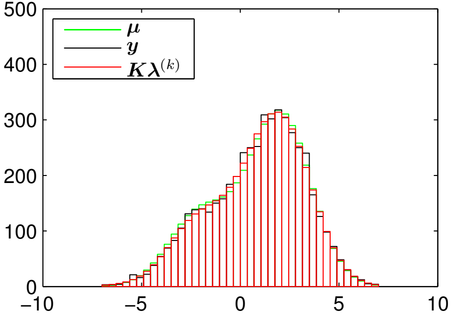


Figure: Smeared histogram

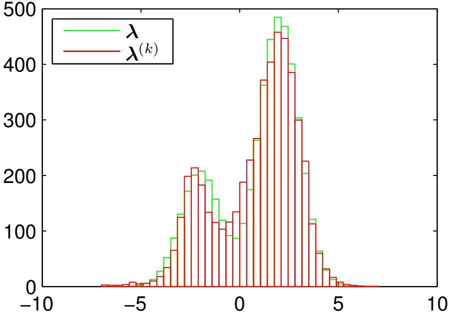


Figure: True histogram

D'Agostini demo, $k = 10000$

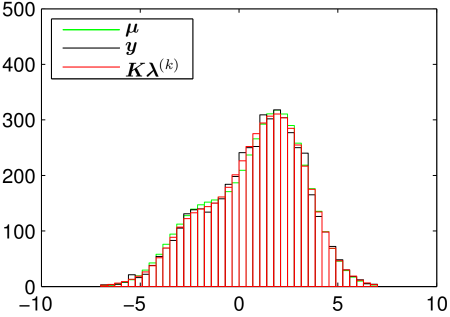


Figure: Smeared histogram

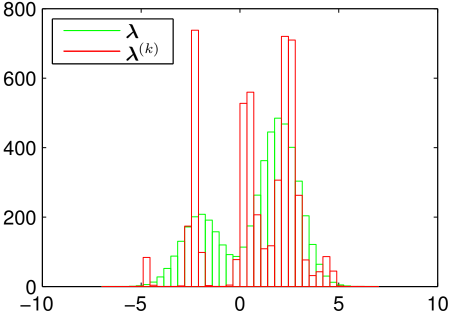


Figure: True histogram

D'Agostini demo, $k = 100000$

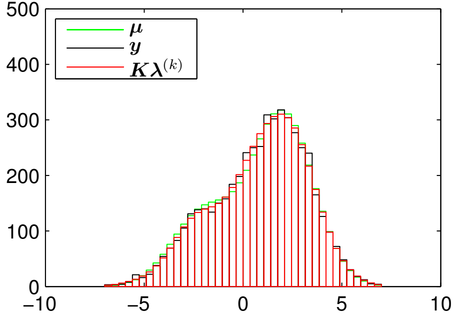


Figure: Smeared histogram

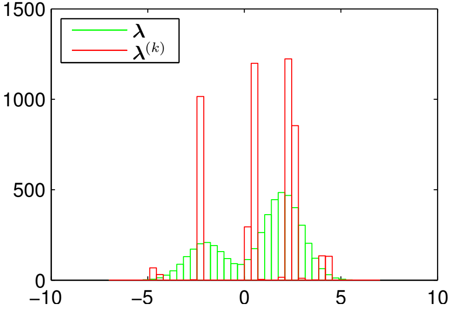


Figure: True histogram

Outline

1. Introduction
2. Basic unfolding methodology
   - Maximum likelihood estimation
   - Regularized frequentist techniques
   - Bayesian unfolding
3. Challenges and opportunities in unfolding
   - Choice of the regularization strength
   - Uncertainty quantification
   - MC dependence of the response matrix
   - Wide-bin unfolding
4. Conclusions

Regularization by early stopping of the EM iteration

- We have seen that unfortunately the MLE itself is often useless
  - Due to the ill-posedness of the problem, it exhibits large, unphysical fluctuations
  - In other words, the likelihood function alone does not contain enough information to constrain the solution
- As the EM iteration proceeds, the solutions will typically first improve but will start to degrade at some point
  - This is because the algorithm will start overfitting to the Poisson fluctuations in $\boldsymbol{y}$
- This behavior can be exploited by stopping the iteration before unphysical features start to appear
  - The number of iterations k now becomes a *regularization parameter* that controls the trade-off between fitting the data and taming unphysical oscillations

D'Agostini demo, $k = 100$

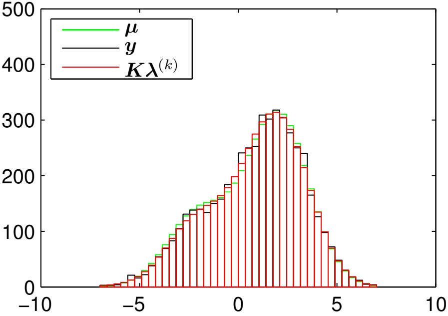


Figure: Smeared histogram

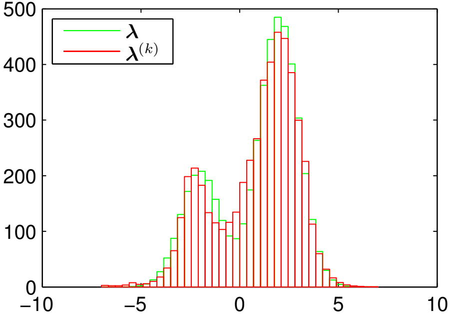


Figure: True histogram

Penalized maximum likelihood estimation

- Early stopping of the EM iteration seems a bit ad-hoc
  - Is there a more principled way of finding good solutions?
- Ideally we would like to find a solution that fits the data but at the same time seems physically plausible
- Let's consider a *penalized maximum likelihood* problem:

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} \; F(\boldsymbol{\lambda}) = \log p(\mathbf{y}|\boldsymbol{\lambda}) - \delta P(\boldsymbol{\lambda})$$

- Here:
  - $P(\boldsymbol{\lambda})$ is a *penalty function* that obtains large values for physically implausible solutions
  - $\delta > 0$ is a *regularization parameter* that controls the balance between maximizing the likelihood and minimizing the penalty
- Typically $P(\boldsymbol{\lambda})$ is a measure of the curvature of the solution
  - I.e., it penalizes for large oscillations

From penalized likelihood to Tikhonov regularization

- To simplify this optimization problem, we use a Gaussian approximation of the Poisson likelihood

$$\mathbf{y}|\boldsymbol{\lambda} \sim \mathrm{Poisson}(\mathbf{K}\boldsymbol{\lambda}) \approx N(\mathbf{K}\boldsymbol{\lambda}, \hat{\mathbf{C}}),$$

  where $\hat{\mathbf{C}} = \mathrm{diag}(\mathbf{y})$

- Hence the objective function becomes:

$$\begin{aligned} F(\boldsymbol{\lambda}) &= \log p(\mathbf{y}|\boldsymbol{\lambda}) - \delta P(\boldsymbol{\lambda}) \\ &= \sum_{i=1}^{n} \left[y_i \log \left(\sum_{j=1}^{p} K_{ij}\lambda_j \right) - \sum_{j=1}^{p} K_{ij}\lambda_j \right] - \delta P(\boldsymbol{\lambda}) + \mathrm{const} \\ &\approx -\frac{1}{2}(\mathbf{y} - \mathbf{K}\boldsymbol{\lambda})^{\mathsf{T}} \hat{\mathbf{C}}^{-1} (\mathbf{y} - \mathbf{K}\boldsymbol{\lambda}) - \delta P(\boldsymbol{\lambda}) + \mathrm{const} \end{aligned}$$

From penalized likelihood to Tikhonov regularization

- Let us drop the positivity constraint and absorb the factor $1/2$ into the penalty to obtain

$$\begin{aligned}\hat{\boldsymbol{\lambda}} &= \operatorname*{arg\,max}_{\boldsymbol{\lambda}\in\mathbb{R}^p} -(\mathbf{y}-\mathbf{K}\boldsymbol{\lambda})^{\mathsf{T}}\hat{\mathbf{C}}^{-1}(\mathbf{y}-\mathbf{K}\boldsymbol{\lambda}) - \delta P(\boldsymbol{\lambda}) \\ &= \operatorname*{arg\,min}_{\boldsymbol{\lambda}\in\mathbb{R}^p} (\mathbf{y}-\mathbf{K}\boldsymbol{\lambda})^{\mathsf{T}}\hat{\mathbf{C}}^{-1}(\mathbf{y}-\mathbf{K}\boldsymbol{\lambda}) + \delta P(\boldsymbol{\lambda})\end{aligned}$$

- We see that we have ended up with a penalized χ^2 problem
- This is typically called *(generalized) Tikhonov regularization*

How to choose the penalty?

- The penalty term should reflect the analyst's a priori understanding of plausible solutions
- Common choices include:
  - Norm of the solution: $P(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|^2$
  - Curvature of the solution: $P(\boldsymbol{\lambda}) = \|\mathbf{L}\boldsymbol{\lambda}\|^2$, where $\mathbf{L}$ is a discretized 2nd derivative operator
  - SVD unfolding (Höcker and Kartvelishvili, 1996):

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

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

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

Explicit form of the Tikhonov estimator

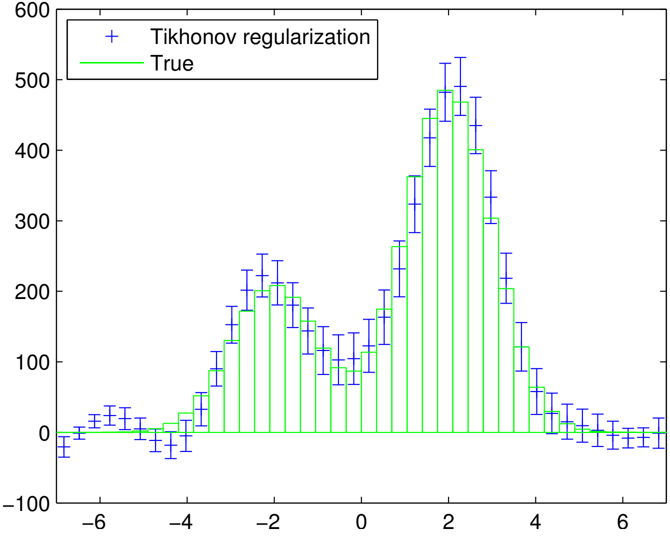
- For all these penalty terms, the Tikhonov regularized point estimator $\hat{\boldsymbol{\lambda}}$ can be written down in closed form
- For instance, consider the problem

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in \mathbb{R}^p}{\arg\min}\ (\mathbf{y} - \mathbf{K}\boldsymbol{\lambda})^{\mathsf{T}} \hat{\mathbf{C}}^{-1} (\mathbf{y} - \mathbf{K}\boldsymbol{\lambda}) + \delta \|\boldsymbol{L}\boldsymbol{\lambda}\|^2$$

- One can easily show (see the backup) that the minimizer is given by

$$\hat{\boldsymbol{\lambda}} = \left(\boldsymbol{K}^{\mathsf{T}} \hat{\boldsymbol{C}}^{-1} \boldsymbol{K} + \delta \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} \right)^{-1} \boldsymbol{K}^{\mathsf{T}} \hat{\boldsymbol{C}}^{-1} \boldsymbol{y}$$

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



Outline

1. Introduction
2. Basic unfolding methodology
   - Maximum likelihood estimation
   - Regularized frequentist techniques
   - Bayesian unfolding
3. Challenges and opportunities in unfolding
   - Choice of the regularization strength
   - Uncertainty quantification
   - MC dependence of the response matrix
   - Wide-bin unfolding
4. Conclusions

Bayesian unfolding

- In Bayesian unfolding, inferences about $\boldsymbol{\lambda}$ are based on the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$
- This is obtained using Bayes' rule:

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int_{\mathbb{R}^p_+} p(\boldsymbol{y}|\boldsymbol{\lambda}')p(\boldsymbol{\lambda}')\,\mathrm{d}\boldsymbol{\lambda}'}, \quad \boldsymbol{\lambda} \in \mathbb{R}^p_+,$$

  where the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$ is the same as earlier and $p(\boldsymbol{\lambda})$ is a prior distribution for $\boldsymbol{\lambda}$
- The most common choices as a point estimator of $\boldsymbol{\lambda}$ are:
  - The *posterior mean*: $\hat{\boldsymbol{\lambda}} = \mathrm{E}[\boldsymbol{\lambda}|\boldsymbol{y}] = \int_{\mathbb{R}^p_+} \boldsymbol{\lambda} p(\boldsymbol{\lambda}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\lambda}$
  - The *maximum a posteriori* (MAP) *estimator*: $\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^p_+} p(\boldsymbol{\lambda}|\boldsymbol{y})$
- The width of the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$ can be used to quantify uncertainty about $\boldsymbol{\lambda}$
  - But note that the interpretation of the resulting Bayesian *credible intervals* is different from frequentist confidence intervals

Regularization using the prior

- In the Bayesian approach, the prior density $p(\boldsymbol{\lambda})$ regularizes the otherwise ill-posed problem
  - It concentrates the probability mass of the posterior on physically plausible solutions
- The prior is typically of the form

$$p(\boldsymbol{\lambda}) \propto \exp\left(-\delta P(\boldsymbol{\lambda})\right), \quad \boldsymbol{\lambda} \in \mathbb{R}_+^p,$$

  where $P(\boldsymbol{\lambda})$ is a function characterizing a priori plausible solutions and $\delta > 0$ is a *hyperparameter* controlling the scale of the prior density
- For example, choosing $P(\boldsymbol{\lambda}) = \|\boldsymbol{L}\boldsymbol{\lambda}\|^2$, where $\boldsymbol{L}$ a discretized 2nd derivative operator, leads to the Gaussian smoothness prior

$$p(\boldsymbol{\lambda}) \propto \exp\left(-\delta \|\boldsymbol{L}\boldsymbol{\lambda}\|^2\right), \quad \boldsymbol{\lambda} \in \mathbb{R}_+^p$$

Connection between Bayesian unfolding and penalized MLE

- Notice that when $p(\boldsymbol{\lambda}) \propto \exp\left(-\delta P(\boldsymbol{\lambda})\right)$, the Bayesian MAP solution coincides with the penalized maximum likelihood estimator:

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}} &= \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} p(\boldsymbol{\lambda}|\boldsymbol{y}) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} \log p(\boldsymbol{\lambda}|\boldsymbol{y}) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} \log p(\boldsymbol{y}|\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} \log p(\boldsymbol{y}|\boldsymbol{\lambda}) - \delta P(\boldsymbol{\lambda}) \\
&= \hat{\boldsymbol{\lambda}}_{\mathrm{PMLE}}
\end{aligned}$$

- So the penalty term $\delta P(\boldsymbol{\lambda})$ can either be interpreted as a Bayesian prior or as a frequentist regularization term
- The Bayesian interpretation has the advantage that we can visualize the prior $p(\boldsymbol{\lambda})$ by, e.g., drawing samples from it

A note about Bayesian computations

- To be able to compute the posterior mean $\mathrm{E}[\boldsymbol{\lambda}|\boldsymbol{y}]$ or form the Bayesian credible intervals, we need to be able to evaluate the posterior

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int_{\mathbb{R}_+^p} p(\boldsymbol{y}|\boldsymbol{\lambda}')p(\boldsymbol{\lambda}')\,\mathrm{d}\boldsymbol{\lambda}'}$$

- But the denominator is an intractable high-dimensional integral...
- Luckily, it turns out that it is possible to *sample* from the posterior without evaluating the denominator
  - The sample mean and sample quantiles can then be used to compute the posterior mean and the credible intervals
- Algorithms that enable this are called Markov chain Monte Carlo (MCMC) samplers and are based on a Markov chain whose equilibrium distribution is the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$
  - The single-component Metropolis–Hastings sampler of Saquib et al. (1998) is particularly well-suited for the unfolding problem and seems to also work well in practice
- As an alternative to MCMC, one could also turn the denominator into a tractable Gaussian integral using the Laplace approximation

Outline

1. Introduction
2. Basic unfolding methodology
   - Maximum likelihood estimation
   - Regularized frequentist techniques
   - Bayesian unfolding
3. Challenges and opportunities in unfolding
   - Choice of the regularization strength
   - Uncertainty quantification
   - MC dependence of the response matrix
   - Wide-bin unfolding
4. Conclusions

Outline

Choice of the regularization strength

- All unfolding methods involve a free parameter controlling the strength of the regularization
  - The parameter δ in Tikhonov regularization and Bayesian unfolding, the number of iterations in D'Agostini
- This parameter is typically difficult to choose using only a priori information
  - But its value usually has a major impact on the unfolded spectrum
- Traditionally many particle physics analyses have chosen the regularization strength using MC studies
  - But this may create an undesired MC bias
- It would be better to choose the regularization strength based on the observed data $\boldsymbol{y}$

Choice of the regularization strength

- Many data-driven methods have been proposed:
  - Cross-validation (Stone, 1974)
  - L-curve (Hansen, 1992)
  - Empirical Bayes estimation (Kuusela and Panaretos, 2015)
  - Goodness-of-fit test in the smeared space (Veklerov and Llacer, 1987)
  - Akaike information criterion (Volobouev, 2015)
  - Minimization of a global correlation coefficient (Schmitt, 2012)
  - ...
- Limited experience about the relative merits of these methods in typical unfolding problems
  - Some evidence that empirical Bayes tends to be more stable than cross-validation (Kuusela, 2016; Wood, 2011)
- Notice that all these are aiming for optimal point estimation
  - Not necessarily optimal for uncertainty quantification!

Outline

Uncertainty quantification

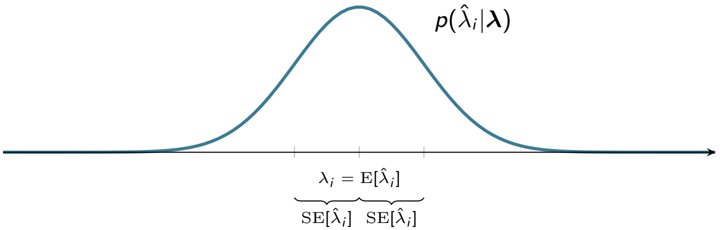
- Proper uncertainty quantification is one of the main challenges in unfolding
- By uncertainty quantification, I mean computing binwise frequentist confidence intervals at $(1-\alpha)$ confidence level:

$$\mathrm{P}_{\boldsymbol{\lambda}}\big(\underline{\lambda}_i(\boldsymbol{y}) \leq \lambda_i \leq \overline{\lambda}_i(\boldsymbol{y})\big) \geq 1-\alpha, \quad \forall i \in 1,\dots,p, \quad \forall \boldsymbol{\lambda} \in \mathbb{R}_+^p$$

- The left-hand side is called the *coverage probability* or simply the *coverage* of the confidence interval $\big[\underline{\lambda}_i(\boldsymbol{y}), \overline{\lambda}_i(\boldsymbol{y})\big]$
- Providing unfolded uncertainties $\big[\underline{\lambda}_i(\boldsymbol{y}), \overline{\lambda}_i(\boldsymbol{y})\big]$ satisfing this inequality is surprisingly tricky!

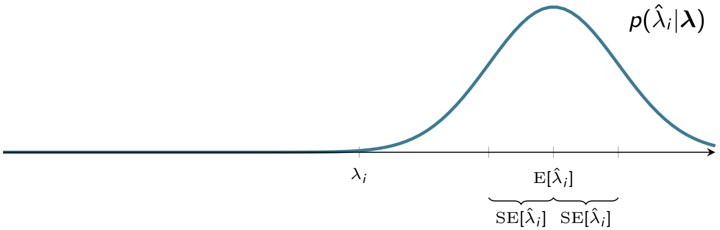
Uncertainty quantification

- Let $\mathrm{SE}[\hat{\lambda}_i]$ be the standard error of $\hat{\lambda}_i$ (i.e., the standard deviation of the sampling distribution of $\hat{\lambda}_i$)
- In many situations, $\hat{\lambda}_i \pm \widehat{\mathrm{SE}}[\hat{\lambda}_i]$ provides a reasonable 68% confidence interval
  - But this is only true when $\hat{\lambda}_i$ is unbiased and approximately Gaussian
- But in regularized unfolding the estimators are always biased!
  - Regularization reduces variance by increasing the bias (*bias-variance trade-off*)
  - Hence the SE confidence intervals may have lousy coverage

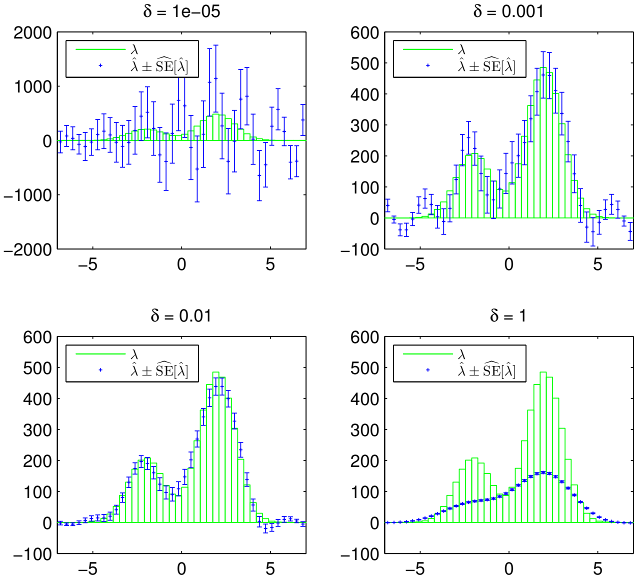


Uncertainty quantification

- Let $\mathrm{SE}[\hat{\lambda}_i]$ be the standard error of $\hat{\lambda}_i$ (i.e., the standard deviation of the sampling distribution of $\hat{\lambda}_i$)
- In many situations, $\hat{\lambda}_i \pm \widehat{\mathrm{SE}}[\hat{\lambda}_i]$ provides a reasonable 68% confidence interval
  - But this is only true when $\hat{\lambda}_i$ is unbiased and approximately Gaussian
- But in regularized unfolding the estimators are always biased!
  - Regularization reduces variance by increasing the bias (*bias-variance trade-off*)
  - Hence the SE confidence intervals may have lousy coverage

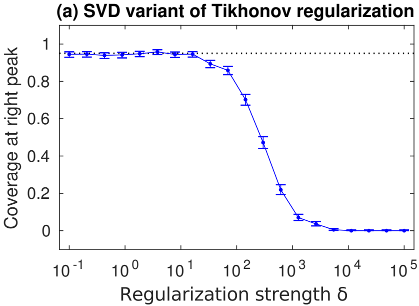


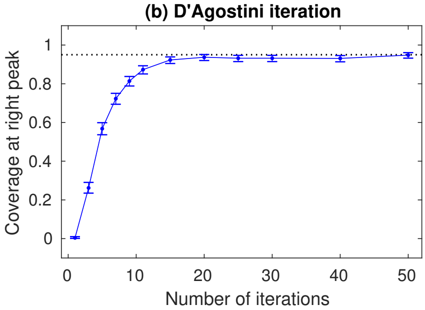
Demonstration with Tikhonov regularization, $P(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|^2$



Uncertainty quantification

- The uncertainties returned by standard software (`RooUnfold`) are estimates of the standard errors computed either using error propagation or resampling
- The coverage of these intervals depends heavily on the regularization strength:





- I have in the past investigated two complementary ways to obtain improved coverage performance:
  1. Debiased intervals (Kuusela and Panaretos, 2015; Kuusela, 2016)
  2. Shape-constrained intervals (Kuusela and Stark, 2016; Kuusela, 2016)

Outline

MC dependence of the response matrix

- The response matrix $\boldsymbol{K}$ is typically estimated using Monte Carlo
- As a result, there are three sources of systematics in $\boldsymbol{K}$:
  1. Finite MC sample size
  2. The matrix depends on the shape of the spectrum within each true bin

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

     where $\{E_i\}_{i=1}^{p}$ and $\{F_i\}_{i=1}^{n}$ are the true and smeared bins, respectively
  3. The smearing of the variable of interest may depend on the MC distribution of some auxiliary variables
     - For example, the energy resolution of jets depends on their pseudorapidity distribution
- Point 2 can be alleviated by making the true bins smaller at the cost of increased ill-posedness

Outline

Unregularized unfolding?

- At the end of the day, *any regularization technique makes unverifiable assumptions about the true spectrum*
  - If these assumptions are not satisfied, the uncertainties will be wrong
- It seems to me that the fundamental problem is that we are asking too hard questions about the true spectrum
  - One simply cannot recover extremely detailed information about f without further outside knowledge
- So the question becomes: What features of f can be recovered based on the smeared data $\boldsymbol{y}$ and how to do this with *honest unregularized* uncertainties?

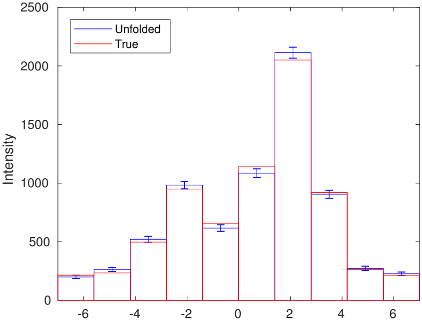
Wide-bin unfolding

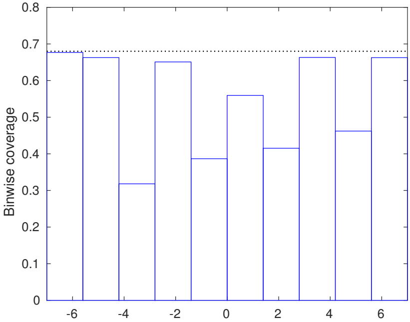
- One functional we should be able to recover without explicit regularization is the integral of f over a *wide* unfolded bin:

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

- 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 and then aggregate into wide bins*
- Let's see how this works using $\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^\dagger \boldsymbol{y}$ and a similar deconvolution setup as before

Wide bins, standard approach, perturbed MC

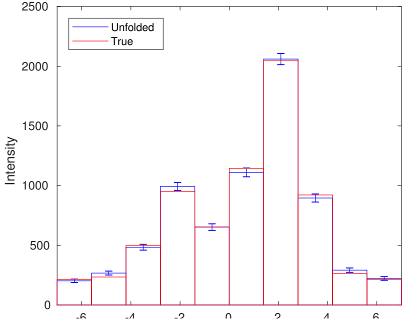


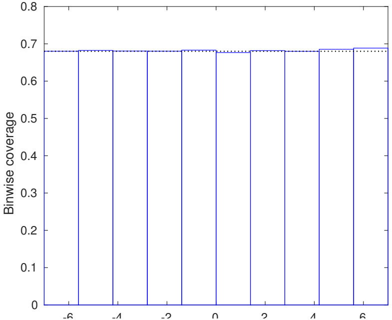


The response matrix $K_{i,j} = \dfrac{\int_{S_i} \int_{T_j} k(s,t) f^{\mathrm{MC}}(t)\,\mathrm{d}t\,\mathrm{d}s}{\int_{T_j} f^{\mathrm{MC}}(t)\,\mathrm{d}t}$ depends on f^{MC}

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

Wide bins, standard approach, correct MC

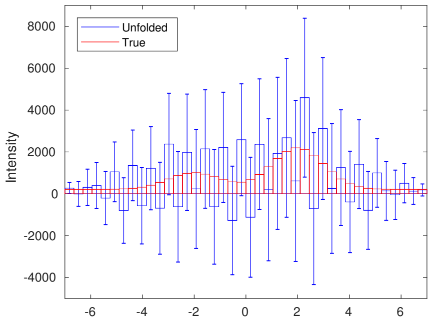


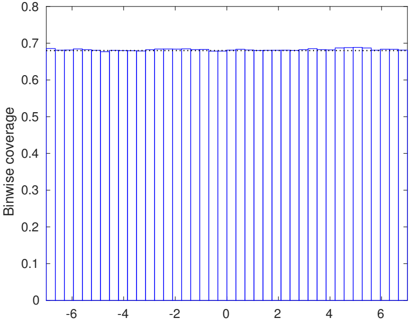


If $f^{\mathrm{MC}} = f$, coverage is correct

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

Fine bins, standard approach, perturbed MC



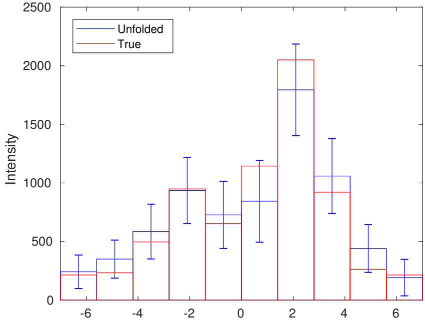


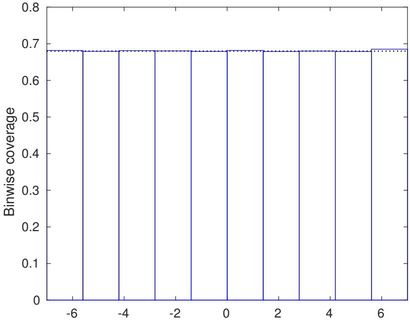
With narrow bins, less dependence on f^{MC} so coverage is correct, but the intervals are very wide[2]

$\Rightarrow$ Let's aggregate these into wide bins, keeping track of the correlations

[2] 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[3]

[3]More unfolded realizations given in the backup.

Outline

Conclusions

- Unfolding is a complex data analysis task with many potential pitfalls
  - It is crucial to understand the ingredients that go into an unfolding procedure
  - Unfolding algorithms should never be used as black boxes!
- All regularized unfolding methods complement the likelihood with additional information about physically plausible solutions
- The most popular techniques are D'Agostini iteration and various flavors of Tikhonov regularization
- Beware when using standard methods that:
  - There is a MC dependence in the smearing matrix and usually also in the regularization
  - The uncertainties do not necessarily provide good coverage performance
  - The regularization parameter has a major impact on the solution and should ideally be chosen in a data-driven way
- It seems that first unfolding with narrow bins followed by aggregation into wide bins provides a way around many of these issues

References I

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.

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.

A. Kondor. Method of convergent weights – An iterative procedure for solving Fredholm's integral equations of the first kind. *Nuclear Instruments and Methods*, 216:177–181, 1983.

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

References II

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. arXiv:1512.00905v3 [stat.AP], submitted, 2016.

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.

H. N. Mülthei and B. Schorr. On an iterative method for a class of integral equations of the first kind. *Mathematical Methods in the Applied Sciences*, 9: 137–168, 1987a.

H. N. Mülthei and B. Schorr. On an iterative method for the unfolding of spectra. *Nuclear Instruments and Methods in Physics Research A*, 257:371–377, 1987b.

References III

H. N. Mülthei and B. Schorr. On properties of the iterative maximum likelihood reconstruction method. *Mathematical Methods in the Applied Sciences*, 11: 331–342, 1989.

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

S. S. Saquib, C. A. Bouman, and K. Sauer. ML parameter estimation for Markov random fields with applications to Bayesian tomography. *IEEE Transactions on Image Processing*, 7(7):1029–1044, 1998.

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. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36: 111–147, 1974.

References IV

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.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73(1):3–36, 2011.

Backup

Maximum likelihood estimation

Theorem (Vardi et al. (1985))

Assume $K_{ij} > 0$ *and* $\mathbf{y} \neq \mathbf{0}$. *Then the following hold for the log-likelihood* $\log p(\mathbf{y}|\boldsymbol{\lambda})$ *of the unfolding problem:*

1. *The log-likelihood has a maximum.*
2. *The log-likelihood is concave and hence all the maxima are global maxima.*
3. *The maximum is unique if and only if the columns of* $\mathbf{K}$ *are linearly independent*

Least squares estimation with the pseudoinverse

- Consider the least squares problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^p} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2,$$

  where $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^n$
- This problem always has a solution, but it may not be unique
- A solution is always given by the Moore–Penrose pseudoinverse of $\boldsymbol{A}$:

$$\hat{\boldsymbol{x}}_{\mathrm{LS}} = \boldsymbol{A}^{\dagger}\boldsymbol{y}$$

- When there are multiple solutions, the pseudoinverse gives the one with the smallest norm
- When $\boldsymbol{A}$ has full column rank, the solution is unique
  - In this special case, the pseudoinverse is given by $\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}}$
  - Hence, the least squares solution is: $\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}$

Finding the Tikhonov regularized solution

- We will now find an explicit form of the Tikhonov regularized estimator
$$\hat{\boldsymbol{\lambda}} = \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in \mathbb{R}^p} (\mathbf{y} - \mathbf{K}\boldsymbol{\lambda})^{\mathsf{T}} \hat{\mathbf{C}}^{-1} (\mathbf{y} - \mathbf{K}\boldsymbol{\lambda}) + \delta \|\boldsymbol{L\lambda}\|^2$$
by rewriting this as a least squares problem
- This approach also easily generalizes to penalty terms involving $\boldsymbol{\lambda}^{\mathrm{MC}}$
- Let us rewrite:
$$\begin{aligned}\hat{\boldsymbol{C}}^{-1} &= \operatorname{diag}\left(\frac{1}{y_1}, \ldots, \frac{1}{y_n}\right) \\ &= \underbrace{\operatorname{diag}\left(\frac{1}{\sqrt{y_1}}, \ldots, \frac{1}{\sqrt{y_n}}\right)}_{:=\boldsymbol{A}} \underbrace{\operatorname{diag}\left(\frac{1}{\sqrt{y_1}}, \ldots, \frac{1}{\sqrt{y_n}}\right)}_{:=\boldsymbol{A}} \\ &= \boldsymbol{AA} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\end{aligned}$$
- Defining $\tilde{\boldsymbol{y}} := \boldsymbol{Ay}$ and $\tilde{\boldsymbol{K}} := \boldsymbol{AK}$, our optimization problem becomes
$$\hat{\boldsymbol{\lambda}} = \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in \mathbb{R}^p} (\tilde{\mathbf{y}} - \tilde{\mathbf{K}}\boldsymbol{\lambda})^{\mathsf{T}} (\tilde{\mathbf{y}} - \tilde{\mathbf{K}}\boldsymbol{\lambda}) + \delta \|\boldsymbol{L\lambda}\|^2$$

Finding the Tikhonov regularized solution

- We can rewrite the objective function as follows:

$$
\begin{aligned}
&(\tilde{\mathbf{y}} - \tilde{\mathbf{K}}\boldsymbol{\lambda})^{\mathsf{T}}(\tilde{\mathbf{y}} - \tilde{\mathbf{K}}\boldsymbol{\lambda}) + \delta\|\boldsymbol{L}\boldsymbol{\lambda}\|^2 \\
&= \|\tilde{\boldsymbol{K}}\boldsymbol{\lambda} - \tilde{\boldsymbol{y}}\|^2 + \|\sqrt{\delta}\boldsymbol{L}\boldsymbol{\lambda}\|^2 \\
&= \left\| \begin{bmatrix} \tilde{\boldsymbol{K}}\boldsymbol{\lambda} - \tilde{\boldsymbol{y}} \\ \sqrt{\delta}\boldsymbol{L}\boldsymbol{\lambda} \end{bmatrix} \right\|^2 \\
&= \left\| \begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix} \boldsymbol{\lambda} - \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \mathbf{0} \end{bmatrix} \right\|^2
\end{aligned}
$$

- Here we recognize a least squares problem, so a minimizer is given by

$$
\hat{\boldsymbol{\lambda}} = \begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix}^{\dagger} \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \mathbf{0} \end{bmatrix}
$$

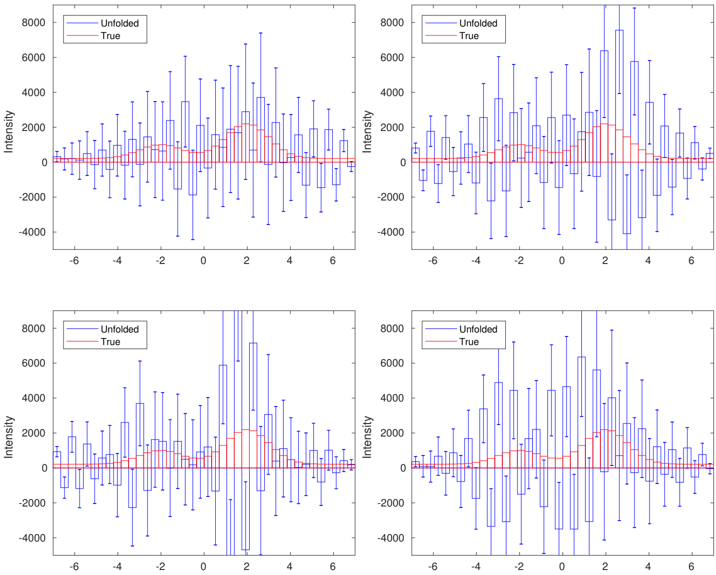
Finding the Tikhonov regularized solution

- Assuming that $\ker(\tilde{\boldsymbol{K}}) \cap \ker(\boldsymbol{L}) = \{\boldsymbol{0}\}$, the minimizer is unique and can be simplified as follows:

$$
\begin{aligned}
\hat{\boldsymbol{\lambda}} &= \begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix}^{\dagger} \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \boldsymbol{0} \end{bmatrix} \\
&= \left(\begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix} \right)^{-1} \begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \boldsymbol{0} \end{bmatrix} \\
&= \left(\begin{bmatrix} \tilde{\boldsymbol{K}}^{\mathsf{T}} & \sqrt{\delta}\boldsymbol{L}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{K}} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix} \right)^{-1} \begin{bmatrix} \tilde{\boldsymbol{K}}^{\mathsf{T}} & \sqrt{\delta}\boldsymbol{L}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \boldsymbol{0} \end{bmatrix} \\
&= \left(\tilde{\boldsymbol{K}}^{\mathsf{T}} \tilde{\boldsymbol{K}} + \delta \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} \right)^{-1} \tilde{\boldsymbol{K}}^{\mathsf{T}} \tilde{\boldsymbol{y}} \\
&= \left(\boldsymbol{K}^{\mathsf{T}} \hat{\boldsymbol{C}}^{-1} \boldsymbol{K} + \delta \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} \right)^{-1} \boldsymbol{K}^{\mathsf{T}} \hat{\boldsymbol{C}}^{-1} \boldsymbol{y}
\end{aligned}
$$

- Hence we have obtained an explicit, closed-form solution for the Tikhonov regularization problem

Fine bins, standard approach, perturbed MC, 4 realizations



Wide bins via fine bins, perturbed MC, 4 realizations