Diagonalizing the Unfolding Problem

– figure out what can be learned and what not –

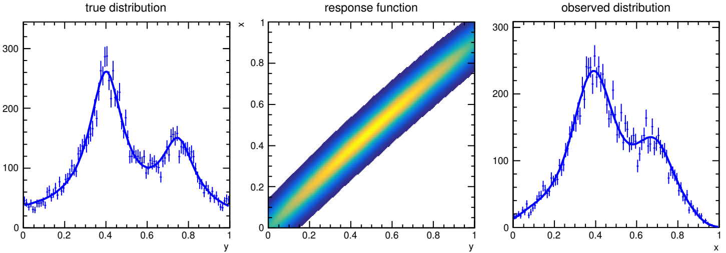
Michael Schmelling / MPI for Nuclear Physics

Disclaimer

This presentation is about ongoing work, i.e. math and story line are still evolving. Only the simplest case of a 1-dim linear unfolding problem is discussed. The talk focuses on results. Mathematical details are given in the backup slides.

Introduction

❖ sketch of a typical 1-dim measurement problem



- given: noisy data that estimate the observable distibution
- known: detector response, i.e. efficiency, biases and smearing
- wanted: an estimate of the true distribution

in the following: linear detector response and counting statistics ➜

The linear model

❖ Fredholm integral equation of 1st kind

$$g(x) = \int dy\, R(x, y)\, f(y)$$

- $f(y)$: true density – unknown
  - integration over the full range allowed for y
  - nature generates values y according to $f(y)$
- $R(x, y)$: response function of the detector – known
  - density in x for fixed y
- $g(x)$: parent distribution of the actual observations – approximately known
  - an experiment records IID measurements x_i, $i = 1, \ldots, N$ drawn from $g(x)$
  - the sample size N is a Poisson distributed random variable

The full story?

❖ real life is more complicated than Fredholm's equation

$$
\begin{aligned}
g(x) = & \int dy\, R_0(x) \\
& + \int dy\, R_1(x, y) f(y) \\
& + \int dy\, dy'\, R_2(x, y, y') f(y) f(y') \\
& + \int dy\, dy'\, dy''\, R_3(x, y, y', y'') f(y) f(y') f(y'') \\
& + \dots
\end{aligned}
$$

back to looking only at the 2nd term ➜

often used: histogram approximation

- integral equation ➜ matrix equation

$$a_k = \sum_{l=1}^{n_b} R_{kl}\, b_l \quad \text{or} \quad a = R\, b$$

with

$$a_k = \int_{\Delta x_k} dx\, g(x)\ ,\ b_l = \int_{\Delta y_l} dy\, f(y) \quad \text{and} \quad R_{kl} = \frac{1}{\Delta y_l} \int_{\Delta x_k} dx \int_{\Delta y_l} dy\, R(x, y)$$

  - a: histogram of the observed distribution
  - b: histogram of the true distribution
  - R: response matrix

- numerically attractive because it requires to handle only finite dimensional matrices
- model dependence from choice of binning

unbinned unfolding ➜

Ansatz: expansion into a complete set of basis functions

❖ starting point: symmetric positive definite operator $T(y, y')$

$$T(y, y') = \int dx\, dx'\, R(x, y)\, R(x', y')\, w(x, x')$$

- appropriate functional form to address the unfolding problem
- any symmetric positive definite weight function $w(x, x')$ is allowed
- simplest choice: $w(x, x') = \delta(x - x')$
- optimum choice for IID data with Poisson distributed sample size:

$$w(x, x') = \frac{\delta(x - x')}{g(x)}$$

  - ► in practice use a Maximum Likelihood estimate $\hat{g}(x)$ instead of $g(x)$
  - ► asymptotic properties are obtained by setting $\hat{g}(x) = g(x)$

❖ properties of the eigenfunctions $b_k(y)$ of $T(y, y')$

$$\int dy'\ T(y, y') b_k(y') = \lambda_k\ b_k(y)$$

- ordered positive eigenvalues $\lambda_k > \lambda_{k+1} > 0\ \forall\ k$
- $w(x, x') \propto 1/N$ ➜ $\lambda_k \propto 1/N$
- orthonormality

$$\int dy\ b_k(y)\ b_l(y) = \delta_{kl}$$

- completeness

$$\sum_{k=0}^{\infty} b_k(y)\ b_k(y') = \delta(y - y')$$

❖ the unfolding problem in the basis $b_k(y)$

$$f(y) = \sum_{k=0}^{\infty} \beta_k b_k(y) \quad \text{and} \quad g(x) = \sum_{k=0}^{\infty} \beta_k\, a_k(x) \quad \text{with} \quad a_k(x) = \int dy\, R(x,y)\, b_k(y)$$

- **note**: from now on specialize to

$$w(x,x') = \frac{\delta(x-x')}{\hat{g}(x)}$$

  - which implies

$$\int dx\, \frac{a_k(x)\, a_l(x)}{\hat{g}(x)} = \lambda_k \delta_{kl} \qquad \text{and it follows} \qquad \int dx\, g(x) \frac{a_k(x)}{\hat{g}(x)} = \lambda_k \beta_k$$

employ this to estimate β_k ➜

❖ use the Law of Large Numbers and that expectation values are integrals

$$\lambda_k \beta_k = \int dx\, g(x) \frac{a_k(x)}{\hat{g}(x)} = N \int dx\, \frac{g(x)}{N} \frac{a_k(x)}{\hat{g}(x)} = N \left\langle \frac{a_k(x)}{\hat{g}(x)} \right\rangle$$

- estimates from a sample of N measurements

$$N \left\langle \frac{a_k(x)}{\hat{g}(x)} \right\rangle \approx \sum_{i=1}^{N} \frac{a_k(x_i)}{\hat{g}(x_i)} = \lambda_k \hat{\beta}_k$$

- covariance matrix of $\hat{\beta}_k$ in the limit $\hat{g}(x) \to g(x)$

$$C_{kl}(\hat{\beta}) = \frac{\delta_{kl}}{\lambda_k}$$

  - ► the expansion coefficients contribute independent information (“diagonalization”)
  - ► the variance grows with higher orders

Information content of the expansion coefficients

- counting experiments: information = number of events = significance^2 $= (n/\sqrt{n})^2$
- significance^2 defines the equivalent number of events in β_k:

$$N_k = \frac{\beta_k^2}{C_{kk}(\beta)} = \beta_k^2 \lambda_k$$

- sum rule in the limit $\hat{g}(x) = g(x)$:

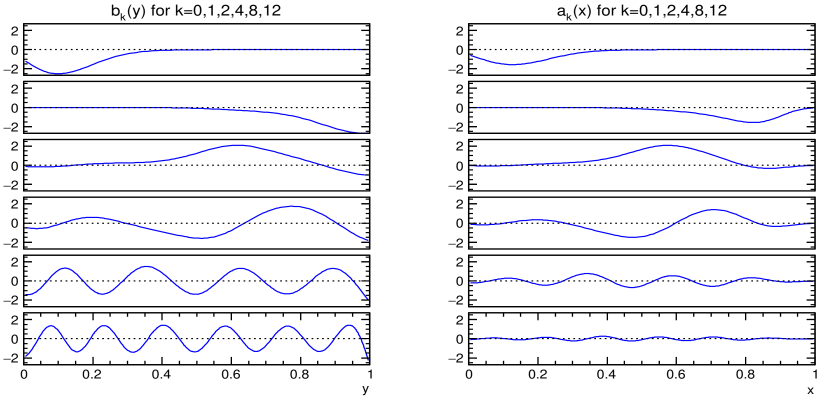
$$\sum_{k=0}^{\infty} N_k = N$$

  - the observed N events are distributed into independent measurements of the β_k
  - the individual shares are proportional to β_k^2 and λ_k
  - only the low-order coefficient get significant shares

numerical studies ➜

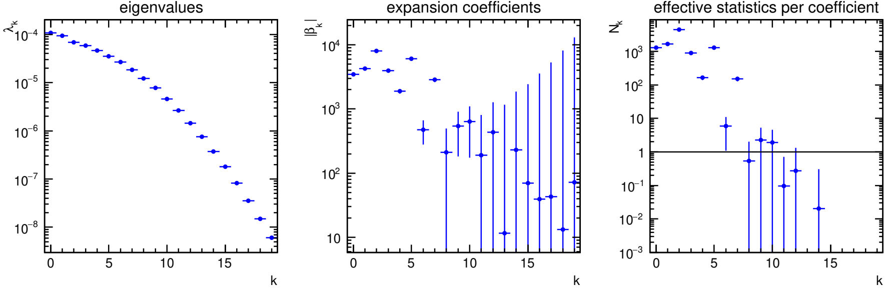
Properties of the introductory example

❖ basis functions for true and observed distribution



❖ coefficients for a low statistics measurement

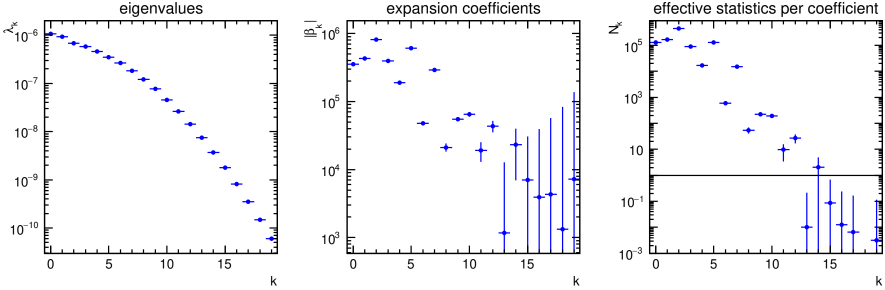
9915 observed events



➜ the experimental information is contained in 10 coefficients

❖ coefficients for a high statistics measurement

999808 observed events



➜ the experimental information is contained in 14 coefficients

Interim summary

- a recipe for an optimal basis $b_k(y)$ to parametrize $f(y)$ has been given
  - the expansion coefficients are independent
  - they are estimated by simple sums of weights over the data
- all information of the data about the true distribution is contained in a small number of leading order coefficients
  - for smooth functions the higher order terms are small $\beta_k \ll 1$
  - higher order terms are suppressed by $\lambda_k \ll 1$
  - the variance of the higher order terms grows with $1/\lambda_k$
- unfolding has been factorized into two problems
  - identifying the available information
  - constructing an estimate $\hat{f}(y)$ of the true distribution $f(y)$

first steps ➜

Truncated expansion with only the well measured coefficients

$$f(y) = \sum_{k=0}^{\infty} \beta_k b_k(y) \quad \rightarrow \quad \hat{f}(y) = \sum_{k=0}^{m} \hat{\beta}_k b_k(y)$$

- relation between $\hat{f}(y)$ and $f(y)$ if $\hat{\beta}_k = \beta_k$

$$\hat{f}(y) = \int dy' \, P(y, y') f(y') \quad \text{with} \quad P(y, y') = \sum_{k=0}^{m} b_k(y) b_k(y')$$

  - ► $P(y, y')$ projects $f(y)$ to the subspace spanned by the first m functions
  - ► completeness of the basis $b_k(y)$ implies $P(y, y') = \delta(y - y')$ for $m \rightarrow \infty$

- different points y and y' are correlated with covariance

$$C(y, y') = \sum_{k=0}^{m} \frac{1}{\lambda_k} b_k(y) b_k(y')$$

Concluding discussion

- every unfolding result $\hat{f}(y)$ is characterized by a posterior response $P(y, y')$
- $\hat{f}(y)$ will be biased even if the coefficients β_k are unbiased
- fluctuations in the β_k can cause unphysical structures $\propto b_k(y)$ in $\hat{f}(y)$
  - estimating $f(y)$ by the truncated expansion may give unsatisfactory results
- the basis $b_k(y)$ is detector specific
  - comparison or combination of experiments not obvious
  - check of theoretical predictions can be done for any basis $b_k(y)$
- more sophisticated regularisation schemes try to inject missing information
  - adjust expansion coefficients within uncertainties
  - invent higher order terms according to some prior knowledge
  - modify posterior response and correlation function
  - but the fundamental limitations of the truncated expansion always apply

Backup

Orthonormality of the basis $b_k(y)$

Application of the eigenvalue equation shows

$$I_{kl} = \int dy\; b_l(y)\; b_k(y) = \frac{1}{\lambda_k} \int dy\; dy'\; b_l(y)\; T(y, y')\; b_k(y') = \frac{\lambda_l}{\lambda_k} \int dy'\; b_l(y)\; b_k(y')$$

and thus

$$I_{kl} = \frac{\lambda_l}{\lambda_k} I_{kl}$$

and it follows

- $\lambda_k \neq \lambda_l$: $I_{kl} = \delta_{kl}$
- $\lambda_k = \lambda_l$: orthogonal linear combinations of $b_k(y)$ and $b_l(y)$ exist

and with appropriate normalization:

$$\int dy\; b_l(y)\; b_k(y) = \delta_{kl}$$

Completeness of the basis $b_k(y)$

Given an arbitrary function $f(y)$ and it's expansion into $b_k(y)$

$$f(y) = \sum_{k=0}^{\infty} \beta_k \, b_k(y) \; ,$$

the coefficients β_k are obtained, using the orthonormality of the $b_k(y)$, by

$$\int dy \; b_k(y) \, f(y) = \int dy \; b_k(y) \sum_{l=0}^{\infty} \beta_l b_l(y) = \sum_{l=0}^{\infty} \beta_l \int dy \; b_k(y) \; b_l(y) = \sum_{l=0}^{\infty} \beta_l \, \delta_{kl} = \beta_k \; .$$

Inserting the result for β_k then yields

$$f(y) = \sum_{k=0}^{\infty} b_k(y) \int dy' \; b_k(y') \, f(y') = \int dy' \, f(y') \sum_{k=0}^{\infty} b_k(y) \; b_k(y')$$

which implies

$$\sum_{k=0}^{\infty} b_k(y) \; b_k(y') = \delta(y - y')$$

Orthogonality of the transformed basis $a_k(x)$

According to the definition of $a_k(x)$ one has

$$\int dx\, dx'\, a_k(x) a_l(x') w(x,x') = \int dy\, dy'\, dx\, dx'\, R(x,y) b_k(y) R(x',y') b_l(y') w(x,x')$$

$$= \int dy\, dy'\, b_k(y)\, T(y,y') b_l(y') = \lambda_l \int dy\, b_k(y)\, b_l(y) = \lambda_l \delta_{kl}$$

which for the special case

$$w(x,x') = \frac{\delta(x-x')}{\hat{g}(x)} \quad \text{simplifies to} \quad \int dx\, \frac{a_k(x)\, a_l(x)}{\hat{g}(x)} = \lambda_k \delta_{kl}$$

and which with one finds

$$\int dx\, g(x)\, \frac{a_k(x)}{\hat{g}(x)} = \int dx \sum_{l=0}^{\infty} \beta_l a_l(x) \frac{a_k(x)}{\hat{g}(x)} = \sum_{l=0}^{\infty} \beta_l \int dx\, \frac{a_l(x) a_k(x)}{\hat{g}(x)} = \lambda_k \beta_k$$

Decomposition of the response matrix

Using the completeness relation and the definition of $a_k(x)$ one finds:

$$
\begin{aligned}
R(x,y) &= \int dy'\, R(x,y')\,\delta(y-y') \\
&= \int dy'\, R(x,y) \sum_{k=0}^{\infty} b_k(y) b_k(y') \\
&= \sum_{k=0}^{\infty} b_k(y) \int dy'\, R(x,y') b_k(y') \\
&= \sum_{k=0}^{\infty} b_k(y) a_k(x) = \sum_{k=0}^{\infty} a_k(x) b_k(y)
\end{aligned}
$$

Covariance matrix of $\hat{\beta}_k$

With the weight functions $u_k(x)$

$$u_k(x) = \frac{1}{\lambda_k} \frac{a_k(x)}{\hat{g}(x)} \quad \text{and thus} \quad \hat{\beta}_k = \sum_{i=1}^{N} u_k(x_i)$$

one finds, for Poisson distributed samples sizes N with $\langle N(N-1)\rangle = \langle N\rangle^2$,

$$\begin{aligned} C_{kl}(\hat{\beta}) &= \langle \hat{\beta}_k \hat{\beta}_l \rangle - \langle \hat{\beta}_k \rangle \langle \hat{\beta}_l \rangle \\ &= \langle N(N-1)\rangle \langle u_k \rangle \langle u_l \rangle + \langle N \rangle \langle u_k u_l \rangle - \langle N \rangle^2 \langle u_k \rangle \langle u_l \rangle \\ &= \langle N \rangle \langle u_k u_l \rangle = \langle N \rangle \int dx \frac{g(x)}{\langle N \rangle} \frac{a_k(x) a_l(x)}{\hat{g}^2(x) \lambda_k \lambda_l} = \int dx \frac{a_k(x) a_l(x)}{\hat{g}(x) \lambda_k \lambda_l} = \frac{\delta_{kl}}{\lambda_k} . \end{aligned}$$

The last two steps follow in the limit $\hat{g}(x) = g(x)$ and from the orthogonality of the $a_k(x)$.

Sum rule

Given the number of equivalent events in β_k

$$N_k = \frac{\beta_k^2}{C_{kk}(\beta)} = \beta_k^2 \lambda_k$$

In the limit $\hat{g}(x) = g(x)$ with

$$\lambda_k \beta_k = \int dx\, g(x) \frac{a_k(x)}{\hat{g}(x)} = \int dx\, a_k(x)$$

one finds

$$\sum_{k=0}^{\infty} N_k = \sum_{k=0}^{\infty} \beta_k (\beta_k \lambda_k) = \sum_{k=0}^{\infty} \beta_k \int dx\, a_k(x) = \int dx \sum_{k=0}^{\infty} \beta_k a_k(x) = \int dx\, g(x) = N\,,$$

i.e. effectively the total data set of N events is split into disjoint measurements of the β_k.

Covariance matrix of unfolded distribution

The covariance between two points y and y' is

$$C(y,y') = \left\langle \hat{f}(y)\hat{f}(y') \right\rangle - \left\langle \hat{f}(y) \right\rangle \left\langle \hat{f}(y') \right\rangle$$

and with

$$f(y) = \sum_{k=0}^{m} \hat{\beta}_k \, b_k(y)$$

one finds

$$\begin{aligned} C(y,y') &= \sum_{k,l=0}^{m} \langle \hat{\beta}_k \hat{\beta}_l \rangle \, b_k(y) b_l(y') - \sum_{k,l=0}^{m} \langle \hat{\beta} \rangle \langle \hat{\beta}_l \rangle \, b_k(y) b_l(y') \\ &= \sum_{k,l=0}^{m} C_{kl}(\hat{\beta}) b_k(y) b_l(y') = \sum_{k,l=0}^{m} \frac{\delta_{kl}}{\lambda_k}] \, b_k(y) b_l(y') = \sum_{k=0}^{m} \frac{1}{\lambda_k} \, b_k(y) b_l(y') \end{aligned}$$