DESY Computing Seminar 23rd June 2008

Unfolding for HEP experiments

Introduction to inverse problems

Volker Blobel – Universität Hamburg

Unfolding is required, due to migration effects, for the measurement of distributions in HEP, where the quantity of interest is often measured indirectly. Unfolding is an linear inverse problem with a coefficient matrix, which is usually ill-conditioned. Techniques of standard linear algebra no longer apply and the numerical treatment becomes more difficult. Available additional information can be used to stabilize the solution, without introducing a significant bias. The use of these regularization methods requires some insight into statistical behavior and mathematical operations.

The measurement of distributions, cross sections . . . is distorted by **migration effects** and **statistical errors** – the transformation of the raw measured data to the measured distribution is called **unfolding**.

All theorems are true,
All models are wrong.
All data are inaccurate.
What are we to do?

L. Smith: Chaos

Unfolding is a complex mathematical operation and requires a good understanding of the detector. Straightforward methods can result in solutions which look chaotic.

---

1. An introductory example — CHARM 1979/1980

Measurement of the differential cross section of

$$\text{inelasticity} \quad y = E_{\text{had}}/E_{\nu,in}$$

in neutral current neutrino interactions, in narrow-band ν-beam with peaks from K and π decay.

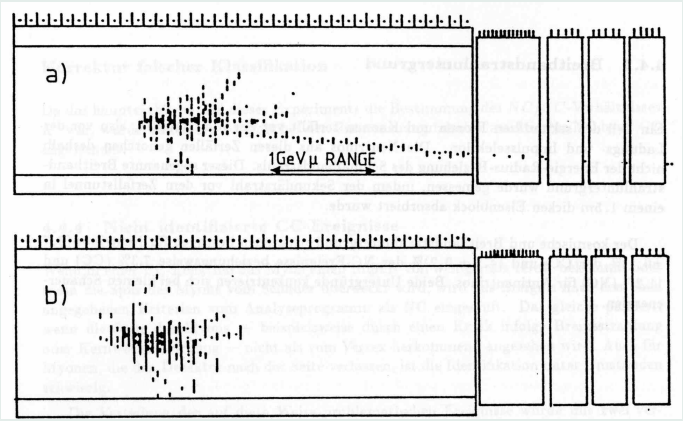
reactions in long calorimeter — measure E_{had} and radius r of vertex — beamflux $\phi(E_\nu, r)$

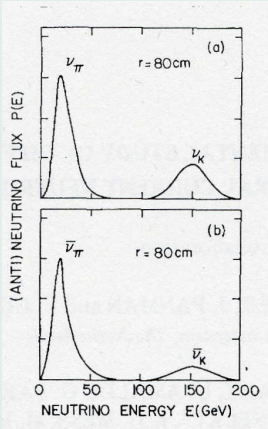
CC: $\nu\, N \to [\text{hadrons}]\, \mu$

$E_{\nu,in} = E_{\text{had}} + E_\mu$

NC: $\nu\, N \to [\text{hadrons}]\, \nu$

$E_{\nu,in} = E_{\text{had}} + ???$





$E_{\nu,in}$ is unmeasurable in NC $\Rightarrow$ use measured radius r of vertex, and the known beamflux $\phi(E_\nu, r)$

Direct and inverse problems

Determination of $\mathrm{d}\sigma/\mathrm{d}y$:

1. **Method:** Most experimental physicists want to measure a physical quantity event-by-event und fill a histogram.
   Strategy: ignore ν from π-decay, use neutrino-energy $E_{\nu,in}$ from flux-peak of ν_K at measured radius –

   and later correct by MC for the wrong assumption.

2. **Method:** It is impossible to reconstruct, event-by-event, the *value of* the inelasticity y for individual events; but the differential cross section $\mathrm{d}\sigma/\mathrm{d}y$ can be reconstructed.
   Strategy:

direct problem (by MC)	bin of $\mathrm{d}\sigma/\mathrm{d}y \Longrightarrow$	predicted distribution in E_{had} and r
inverse problem (by unfolding)	$\mathrm{d}\sigma/\mathrm{d}y \Longleftarrow$	measured distribution in E_{had} and r

In this case unfolding is the transformation of the measured two-dimensional distribution of (E_{had}, r) to the one-dimensional distribution $\mathrm{d}\sigma/\mathrm{d}y$, using the response matrix determined by MC.

In a real experiment there is a set of observations, consisting of noisy measurements, which may be biased, and are fundamentally different fom the noise-free mathematical and physical predictions - statistical and systematic errors!

Reality is that which,
when you stop believing in it,
doesn't go away.

P.K. Dick

Two examples for systematic problems/errors in the Charm experiment

Ratio of ν_π to ν_K in beamflux:

if this is wrong, the whole result will become distorted.
Is it possible to make an independent check?

Hadronic energy calibration:

if this is wrong e.g. to high by 3 %, the result is a spike at high y.
An almost constant shape is expected at high y: use this information to improve the calibration?
If yes $\Rightarrow$ no measurement at high y.

A correct determination of the response matrix A (e.g. by Monte Carlo) is essential.

This will require a detailed and precise knowledge of the detector behaviour – the . . .

Art of Physics Analysis

. . . that is not treated in textbooks.

2. Inverse problems and unfolding

The transition input $\boldsymbol{x} \Rightarrow$ output $\boldsymbol{y}$ in a system is, in general form

$$\int_{\Omega} \text{system} \times \text{input} \; \mathrm{d}\Omega = \text{output} \; ,$$

expressed by the Fredholm integral equation of the first kind

$$\int_{\Omega} A(y,x) f(x) \; \mathrm{d}x = g(y) \; ,$$

and after discretization with bins in x, y and $\boldsymbol{A}$, expressed by a linear equation

$$\boxed{\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{y}}$$

The inverse problem is to reconstruct input $\boldsymbol{x}$ from a measured output $\boldsymbol{y}$.

The $\cong$ symbol is used (instead $=$) in the *boxed* equation, because $\boldsymbol{y}$ has measurement errors (or is Poisson distributed) and $m \geq n$:

$f(x) =$	true distribution of x	$\boldsymbol{x} =$	n-histogram of true variable x
$g(y) =$	measured distribution	$\boldsymbol{y} =$	m-histogram of measured variable y
$A(y,x) =$	Kernel, response function	$\boldsymbol{A} =$	$m \times n$ response matrix

Inverse problem are called ill-posed:

"a small perturbation of the data can cause an arbitrary large perturbation of the solution!"

The elements of the matrix $\boldsymbol{A}$ are determined from a Monte Carlo event sample; they do not depend on the distribution $f(x)$, assumed in the Monte Carlo simulation.

Equation $\boldsymbol{Ax} \cong \boldsymbol{y}$ (Measured histogram $\boldsymbol{y}$ is m-vector, result histogram $\boldsymbol{x}$ is n-vector)

Matrix $\boldsymbol{A}$ is m-by-n matrix in detail:

$$\begin{aligned}
A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \ldots\ A_{1n}x_n &\cong y_1 \\
A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \ldots\ A_{2n}x_n &\cong y_2 \\
A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \ldots\ A_{3n}x_n &\cong y_3 \\
&\ldots \\
A_{m1}x_1 + A_{m2}x_2 + A_{m3}x_3 + \ldots A_{mn}x_n &\cong y_m
\end{aligned}$$

where $m \geq n$ is assumed.

The matrix elements A_{ij} can be interpreted in terms of probability:

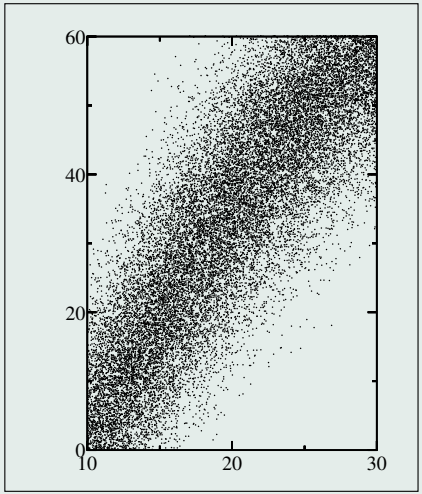
- A_{ij} is the probability for an event originating from bin j of x, to be observed in bin i of y.
- Acceptance probability for events from x-bin j (with $j = 1 \ldots n$) is

$$P_j = \sum_{i=1}^{m} A_{ij} \qquad \text{(column sum of } \boldsymbol{A}\text{)} \ ,$$

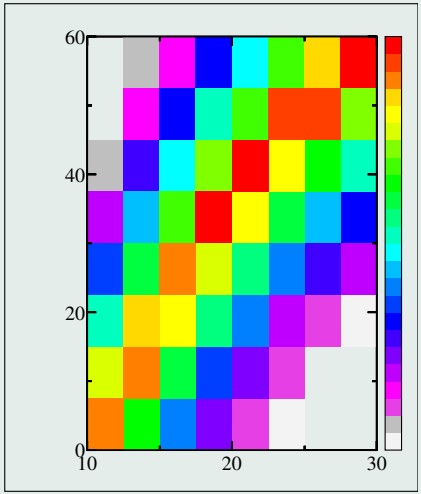
  which often is < 1.

The response matrix $\boldsymbol{A}$ is generated from Monte Carlo xy-pairs.

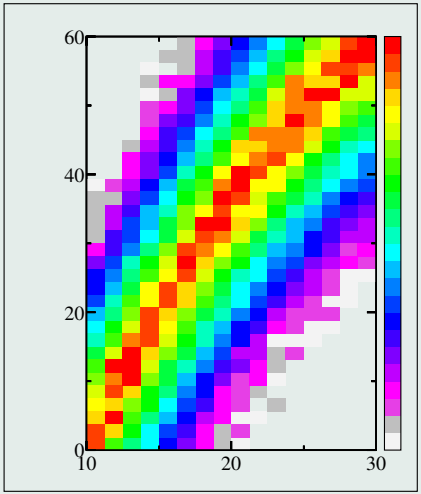
The resolution is deteriorated, if too few bins are used (m too small).



scatter plot x



raw 8×8 bins x



fine 16×32 bins x

Example: plots of x versus y, with small non-linearity in $y = y(x)$ . . .

use number of y-bins $m \gtrsim 2n$

Never use $n = m$ with identical bins! – (“inverse crime”: . . . the numerical methods contain features that effectively render the inverse problem less ill-posed than it actually is, thus yielding unrealistically optimistic results.)

Inverse problems in other fields

Linear inversion: (measurement) $\boldsymbol{y} \Rightarrow \boldsymbol{x}$ (unfolded distribution)

There is an extensive literature (see below) about unfolding, which is of great importance in many different fields:

- tomography in medicine, archaeology, biology, geophysics, oceanography, materials science, astrophysics ... using several different physical phenomena including
  X-rays, gamma rays, positron-electron annihilation reactions, nuclear magnetic resonance, ultrasound, electrons, ions ...;
- image restauration, picture deblurring;
- particle (e.g. neutron) spectra reconstruction from measured pulse-height distributions.

The dimension parameters m and n can be very large.
The covariance matrix $\boldsymbol{V}_x$ is in general *not* required.

Literature:
Per Christian Hansen, Rank-Deficient and Discrete Ill-posed Problems, Siam (1998)
Jari Kaipio and Erkki Somersalo, Statistical and Computational Inverse problems, Springer (2005)

You can get it wrong and still you think it's all right
JOHN LENNON AND PAUL MCCARTNEY from
We can work it out

The requirements for the solution of the unfolding equation

$$\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{y}$$

- Complete usage of all available measurement information, including the (usually diagonal) covariance matrix $\boldsymbol{V}_y$ of measured vector $\boldsymbol{y}$ with least squares (or Poisson likelihood).
- Unfolding should introduce **no bias**, with respect to a particular model of the physical process and Monte Carlo simulation.
- Unfolding should introduce **no or only a small bias**, with respect to general requirements of the solution (a general requirement is e.g. *smoothness* of the solution).
- Complete covariance matrix $\boldsymbol{V}_x$ of solution (and weight matrix $\boldsymbol{W}_x$ = inverse covariance matrix) to be used in subsequent fits of parametrizations to the data.
  For a linear inverse problem a **linear solution** and **standard error propagation** with a matrix $\boldsymbol{A}^{\#}$:

$$\boldsymbol{x} = \boldsymbol{A}^{\#}\boldsymbol{y} \qquad\qquad \boldsymbol{V}_x = \boldsymbol{A}^{\#}\,\boldsymbol{V}_y\boldsymbol{A}^{\#\mathrm{T}}$$

- Solution should have small correlations between different $\boldsymbol{x}$-bins. Correlation coefficients ρ_{ij} up to ± 0.5 are acceptable, values $|\pm\rho_{ij}| > 0.9$ should be avoided, i.e. the matrix $\boldsymbol{V}_x$ should be *almost* diagonal.

Optimal solution in the least squares sense of the equation $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{y}$ is defined by the requirement

$$F(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 = \text{minimum}$$

In the following the covariance matrix $\boldsymbol{V}_y$ is inserted:

$$F(\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}}\, \boldsymbol{V}_y^{-1}\, (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) = \text{minimum}$$

In the normal-equation formalism for the solution the matrix equation

$$\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A}\right)\boldsymbol{x} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{y}\right)$$
$$\boldsymbol{C}\boldsymbol{x} = \boldsymbol{b} \qquad \boldsymbol{C} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A}\right) \qquad \boldsymbol{b} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\right)\boldsymbol{y}$$

has to be solved. The solution vector $\boldsymbol{x}$ is a linear transformation of the measurement vector $\boldsymbol{y}$:

$$\boldsymbol{x} = \boldsymbol{A}^{\#}\boldsymbol{y} \qquad \boldsymbol{A}^{\#} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}$$

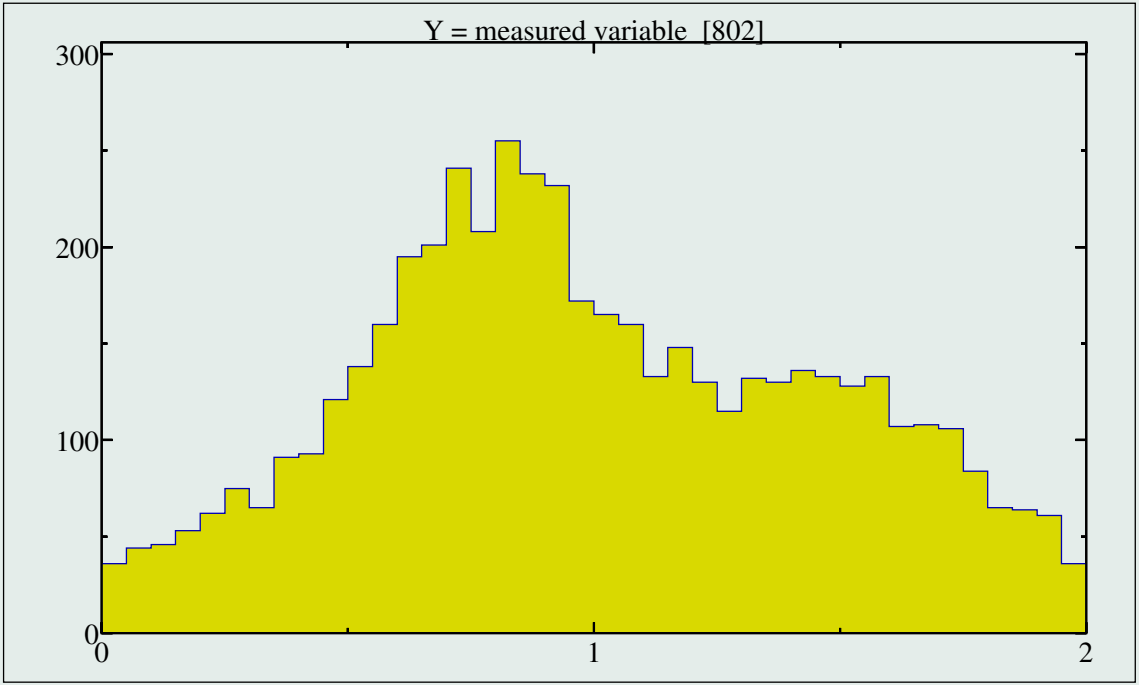
The result of standard error propagation is simple:

$$\boldsymbol{V}_x = \boldsymbol{A}^{\#}\, \boldsymbol{V}_y \boldsymbol{A}^{\#\mathrm{T}} = \boldsymbol{C}^{-1}$$

This error propagation is possible only if the matrix $\boldsymbol{A}^{\#}$ exists in the used numerical method!

The measured distribution

Measured histogram in 40 bins ...



... reconstruction by unfolding will be done in 16 bins, with σ = bin-width.

Properties of the LS solution

The potential bias and the accuracy of a method should be checked.

(1) Check of a potential bias in the solution:

$$\text{estimator} \qquad \widehat{\boldsymbol{x}} = \boldsymbol{A}^{\#}\boldsymbol{y} \quad \text{with} \quad E\left[\boldsymbol{y}\right] = \boldsymbol{A}\boldsymbol{x}_{\text{exact}}$$
$$E\left[\widehat{\boldsymbol{x}}\right] = \boldsymbol{A}^{\#}E\left[\boldsymbol{y}\right] = \left(\boldsymbol{A}^{\#}\boldsymbol{A}\right)\boldsymbol{x}_{\text{exact}} = \boldsymbol{x}_{\text{exact}}$$

$\Rightarrow$ the estimator $\widehat{\boldsymbol{x}}$ is unbiased, because: resolution matrix $\boldsymbol{\Xi} \equiv \boldsymbol{A}^{\#}\boldsymbol{A} = \mathbf{1}$.

(2) Variance: Lower bound of the variance is given by the Rao-Cramér-Frechet (RCF) inequality. The covariance matrix $\boldsymbol{V}_x$ is equal to the lower bound:
$\Rightarrow$ the estimator $\widehat{\boldsymbol{x}}$ has the smallest possible variance for an estimator *with zero bias*.

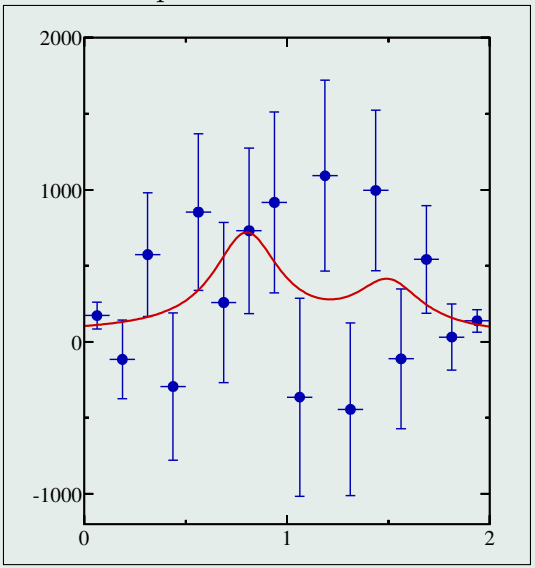
General statement by the Gauss-Markov theorem: the least square estimate is unbiased and efficient.

But: the result $\widehat{\boldsymbol{x}}$ will often show large, unacceptable fluctuations!

The fluctuations are not caused by inaccurate matrix elements (from Monte Carlo), but are inherent in the problem, i.e. the response matrix $\boldsymbol{A}$ and its "smoothing" properties.

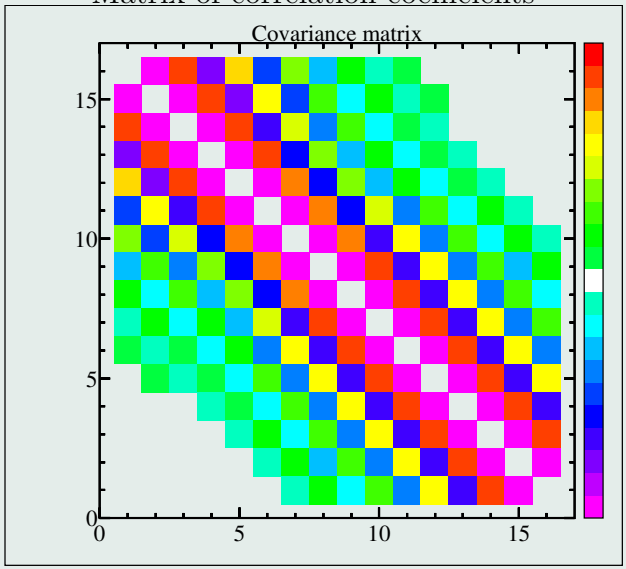
Example of unfolding problem with σ = bin size (Gaussian resolution)

Data points with error bars



True curve $f(x)$ is shown in red.

Matrix of correlation coefficients



$-1 \leq \rho_{ij} \leq +1$

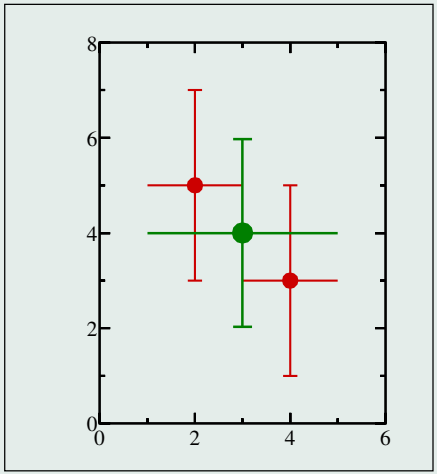
The wonderful world of correlations

Average of two *correlated* numbers d_1 and d_2 (assuming $\sigma_1 = \sigma_2$) with positive/negative correlation:

$$\text{average} \quad \overline{\boldsymbol{d}} = \tfrac{1}{2}(d_1 + d_2)$$
$$\boldsymbol{V}_{\overline{d}} = \tfrac{1}{2}(1 + \rho_{12})\,\sigma^2$$

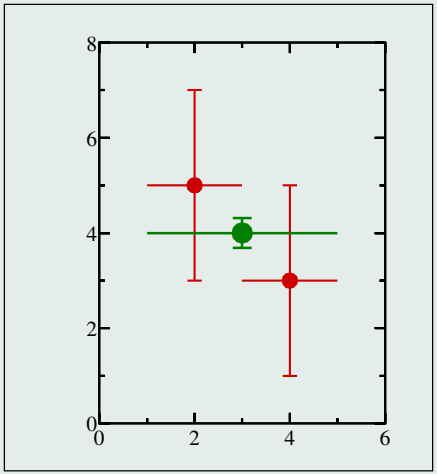
$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$\rho_{12} = +0.95$



Averaged value has almost the same error as each single data value (0.987σ).

$\rho_{12} = -0.95$



Averaged value has much smaller error than each single data value (0.158σ).

Matrix diagonalization

Study transformation properties of symmetric n-by-n matrix $\boldsymbol{C} \quad = (\boldsymbol{A}\boldsymbol{V}_y^{-1}\boldsymbol{A}^{\mathrm{T}})$

by decomposition in the form

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} \qquad \boldsymbol{\Lambda} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{U} \qquad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{I} \qquad \boldsymbol{C}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{\mathrm{T}}$$

with a diagonal matrix $\boldsymbol{\Lambda}$.

All matrices in the decomposition are n-by-n matrices:

$$\begin{pmatrix} & \\ & \boldsymbol{C} \\ & \end{pmatrix} = \begin{pmatrix} & \\ & \boldsymbol{U} \\ & \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \cdot \begin{pmatrix} & \\ & \boldsymbol{U}^{\mathrm{T}} \\ & \end{pmatrix}$$

Eigenvalues $\lambda_j \geq 0$ are in decreasing order, with $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$.

The orthogonal matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ is an array of column vectors = eigenvectors $\boldsymbol{u}_j$ of the matrix $\boldsymbol{C}$.

If the inverse $\boldsymbol{C}^{-1}$ exists (i.e. all $\lambda_j > 0$):

$$\boldsymbol{x} = \boldsymbol{C}^{-1}\boldsymbol{b} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2} \quad \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\,\boldsymbol{b} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{c} \qquad \text{with coefficients} \quad \boldsymbol{c} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b}\,,$$

i.e. the solution is expressed by a vector $\boldsymbol{c}$ of Fourier coefficients.

The covariance matrix of the vector of Fourier coefficients $\boldsymbol{c}$ is simply the unit matrix, because

$$\boldsymbol{V}_c = \left(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{V}_b\left(\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\right) = \left(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{C}\left(\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\right) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{1}\,.$$

The solution can be written with a sum in the form

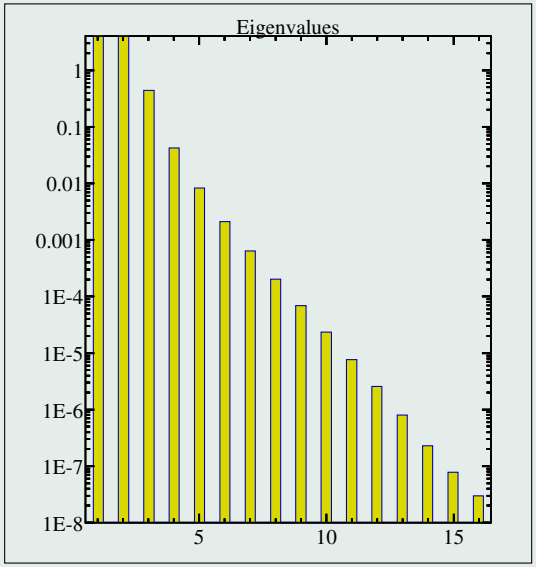
$$\text{Solution:} \qquad \boldsymbol{x} = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} c_j \boldsymbol{u}_j \qquad \text{with} \quad c_j = \frac{1}{\sqrt{\lambda_j}}\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{u}_j\right) \qquad \boldsymbol{V}_x = \sum_{j=1}^{n} \frac{1}{\lambda_j}\boldsymbol{u}_j\boldsymbol{u}_j^{\mathrm{T}}$$

Fourier coefficients c_j, which are insignificant (compatible with zero), should follow a normal distribution $N(0,1)$; this allows simple statistical tests for significance.

The insignificant coefficients can make a huge and dominating contribution to the solution $\boldsymbol{x}$, if the eigenvalues λ_j are small. Diagonalization allows to determine the degree of freedom = number of significant coefficients

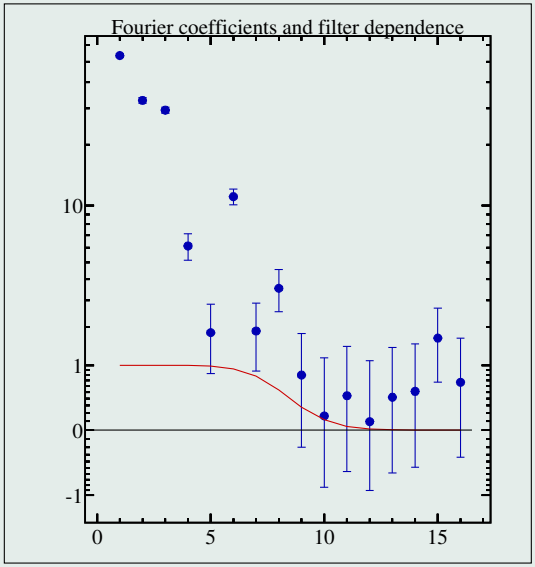
Eigenvalue spectrum and Fourier coefficients

Eigenvalues $= \lambda_j$



The eigenvalues decrease by ≈ 8 orders of magnitude, due to limited resolution.

Fourier coefficients $= c_j$



Note: the c_j are independent, and all have error $1 \Rightarrow$ only ten coefficients are significant.

Truncation: The truncated solution can be written with a **sum up to** $k = 10$ only (instead of n) ... the noisy contributions are removed from the result, without the introduction of a bias, ... but the rank of $\boldsymbol{V}_x$ is only k ($< n$) ($\Rightarrow \boldsymbol{V}_x$ is singular).

4. Regularization methods

Key idea: incorporate certain a-priori assumptions about the size and/or smoothness of the solution!
$\Rightarrow$ control the *norm of the residuals* and, simultaneously, the *norm of the solution* $\boldsymbol{x}$.

$$\textit{Thikhonov-Phillips:} \quad F_\tau(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \tau \, \|\boldsymbol{L}\boldsymbol{x}\|^2 = \text{minimum} \qquad \tau > 0 \qquad \boldsymbol{L} = \boldsymbol{1}$$

$$\text{Matrix equation:} \quad \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A} + \tau \cdot \boldsymbol{1}\right) \quad \boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{y} \qquad \text{to be solved}$$

$$\left(\boldsymbol{C} + \tau \cdot \boldsymbol{1}\right) \quad \boldsymbol{x} = \boldsymbol{b}$$

$$\text{(diagonalization of } \boldsymbol{C}) \qquad \boldsymbol{U}\left(\boldsymbol{\Lambda} + \tau \cdot \boldsymbol{1}\right)\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}$$

$$\text{Fourier coefficients} \quad \boldsymbol{c} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b} \qquad \text{filter factor } f_j = \frac{\lambda_j}{\lambda_j + \tau} = 1 \ \ldots \ 0.5 \ \ldots \ 0$$

$$\text{Solution:} \quad \boldsymbol{x} = \sum_{j=1}^{n} \frac{f_j}{\sqrt{\lambda_j}} c_j \boldsymbol{u}_j \qquad \text{with} \quad c_j = \frac{1}{\sqrt{\lambda_j}} \left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{u}_j\right) \quad (\pm 1)$$

The result $\boldsymbol{x}$ is expressed as a superposition of eigenvectors $\boldsymbol{u}_j$, each weighted with the Fourier coefficient c_j and $1/\sqrt{\lambda_j}$ (!); the filter factor f_j reduces the effect of insignificant contributions, without introducing a bias.

Second derivative

Key idea: incorporate certain a-priori assumptions about the size and/or smoothness of the solution!
$\Rightarrow$ control the *norm of the residuals* and, simultaneously, the *norm of second derivative* $\boldsymbol{Lx}$.

$$F_\tau(\boldsymbol{x}) = \|\boldsymbol{Ax} - \boldsymbol{y}\|^2 + \tau \|\boldsymbol{Lx}\|^2 = \text{minimum} \qquad \tau > 0$$

Approximate second derivative:

$$x_i'' \;\propto\; x_{i-1} - 2x_i + x_{i+1} \qquad \boldsymbol{L} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}$$

Matrix equation:
$$\begin{aligned} \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A} + \tau \cdot \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right) \quad & \boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{y} \qquad \text{to be solved} \\ \left(\boldsymbol{C} + \tau \cdot \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right) \quad & \boldsymbol{x} = \boldsymbol{b} \\ \boldsymbol{R}\left(\mathbf{1} + \tau \cdot \quad \boldsymbol{S}\right)\boldsymbol{R}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b} \qquad & \text{with} \quad \boldsymbol{R} = \boldsymbol{U}_1\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}_2 \end{aligned}$$

requires simultaneous diagonalization of two symmetric matrices $\boldsymbol{C}$ and $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ ($\boldsymbol{S}$ is diagonal).

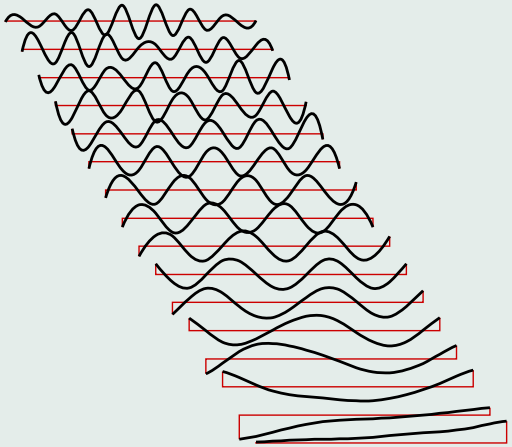
$$\text{Fourier coefficients} \quad \boldsymbol{c} = \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{b} \qquad \text{filter factor } f_i = \frac{\lambda_j}{\lambda_j + \tau} \qquad \text{with} \quad \lambda = S_{jj}^{-1}$$

Often better than standard regularization ($\boldsymbol{L} = \mathbf{1}$), but depends on order of bins (2-dim solution?)!

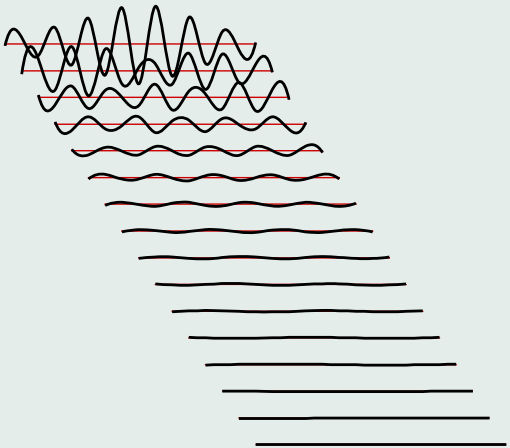
Eigenvectors

The result $\boldsymbol{x}$ is expressed as a superposition of eigenvectors $\boldsymbol{u}_j$, each weighted with the Fourier coefficient c_j and $1/\sqrt{\lambda_j}$ (!); the filter factor f_j reduces the effect of insignificant contributions.

eigenvectors,
(normalized)



eigenvectors,
scaled by eigenvalue factor $1/\sqrt{\lambda_j}$ (!)



Regularization parameter I

Discrepancy principle (Morozov): choose τ such that $\|\boldsymbol{A}\boldsymbol{x}_{\text{reg}} - \boldsymbol{y}\|_2^2 = n_{df}$

L-curve method: plot $\|\boldsymbol{L}\boldsymbol{x}_{\text{reg}}\|_2$ versus $\|\boldsymbol{A}\boldsymbol{x}_{\text{reg}} - \boldsymbol{y}\|_2$ for a set of τ-values $\Rightarrow$ dependence has shape of an L with a distinct corner: optimal value of τ.

Effective number of degrees of freedom: (RUN) determine n_{df} from spectrum of Fourier coefficients c_j, and determine τ such that sum of filter factors

$$\sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} = n_{df}$$

Minimum of global correlation: minimum mean value of global correlation coefficients

Definition: $\rho_j = \sqrt{1 - \left[\left(\boldsymbol{V}_x\right)_{jj} \cdot \left(\boldsymbol{V}_x^{-1}\right)_{jj} \right]^{-1}}$ with $0 \leq \rho_j \leq 1$

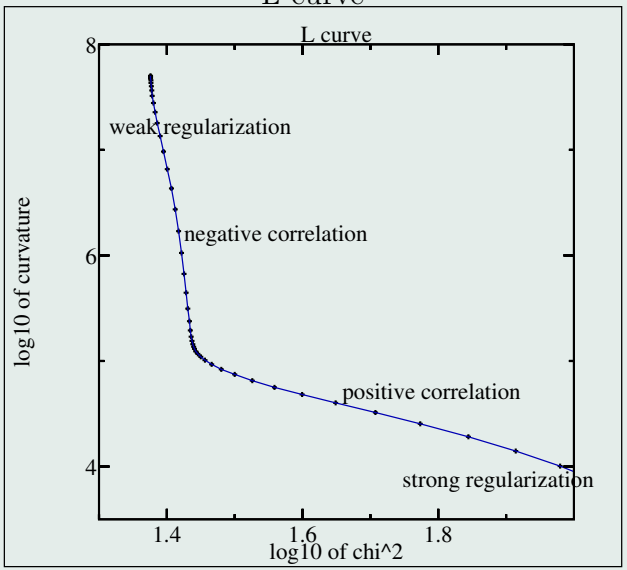
The global correlation coefficient is a measure of the total amount of correlation between element j of $\boldsymbol{x}$ and all other elements.

The arithmetic and the geometric mean of all n global correlation coefficients is determined for a large range of τ-values: the τ-value with the smallest mean value is accepted.

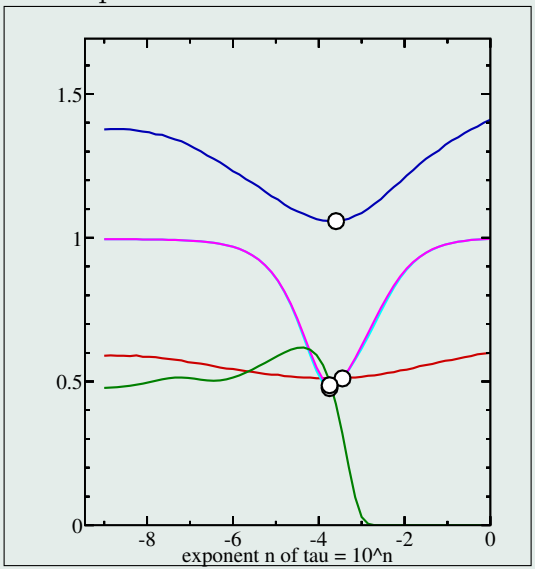
. . . seems to be the best method!

Regularization parameter II

L-curve



Dependence as a function of τ



optimal regularization in region of largest curvature

circle=minimum

magenta/cyan=mean global correlation

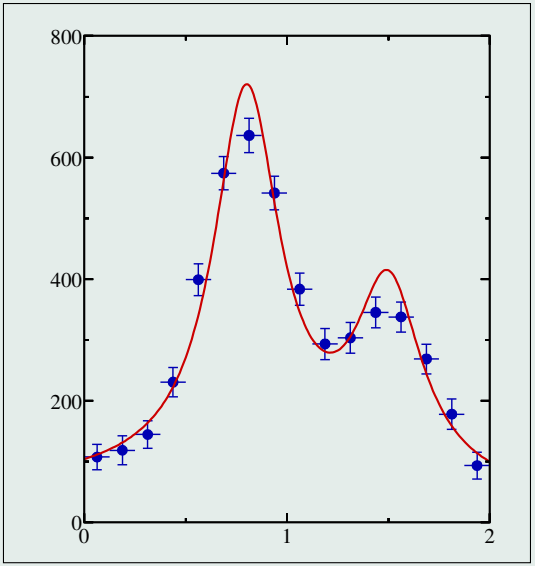
green=χ^2 probability

red=average probability

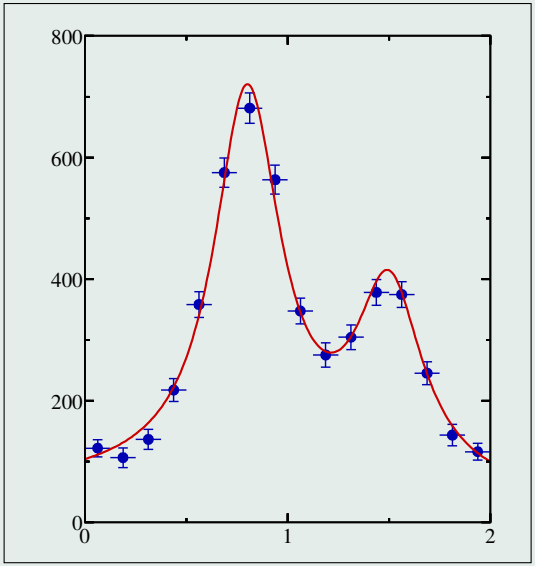
blue=relative sigma

Regularization parameter τ taken from minimum of mean of global correlation coefficients:

$\sigma =$ bin-width $\qquad\qquad$ $\sigma = \frac{1}{2} \times$ bin-width



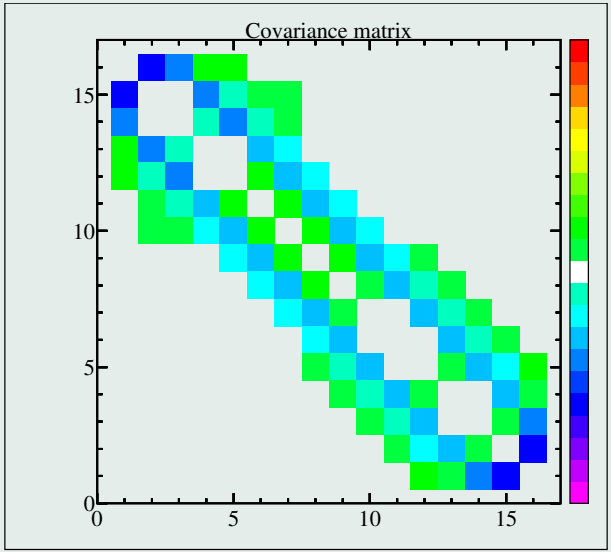
Small bias at the peaks.



Reduced bias at the peaks.

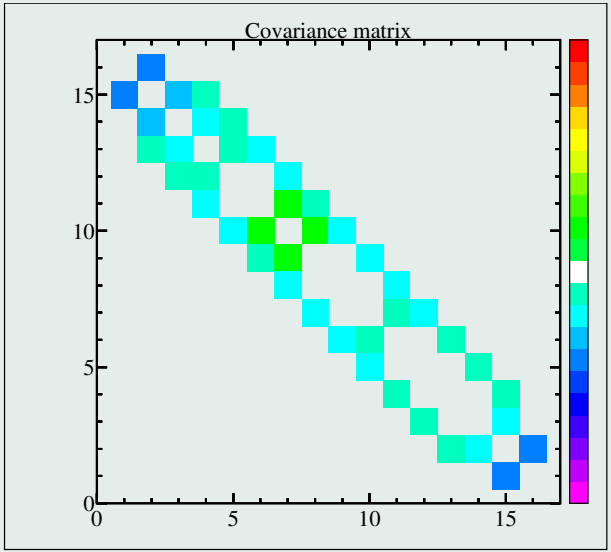
Correlation matrices

$\sigma =$ bin-width



Correlations $\rho_{j,j+1} \approx +0.1$ and $\rho_{j,j+2} \approx -0.25$

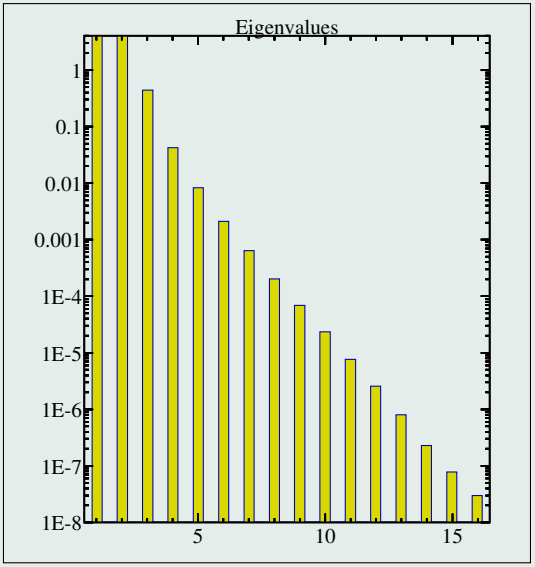
$\sigma = \frac{1}{2} \times$ bin-width



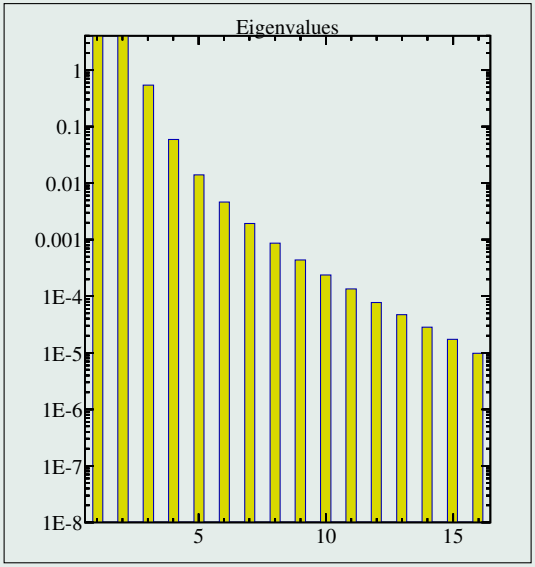
Correlations $\rho_{j,j+1} \approx \pm 0.08$ and $\rho_{j,j+2} \approx -0.1$

Eigenvalue spectrum

$\sigma =$ bin-width



$\sigma = \frac{1}{2} \times$ bin-width



Eigenvalues decrease by a factor of $\approx 10^{-8}$ on the left, and only by $\approx 10^{-6}$ for the better resolution.

5. Iterative methods

Often a method is used where a single computational step is not sufficient to get the solution.
Iterative methods with many computational steps are extremely popular in HEP for problems like unfolding, calibration, alignment ... (need start vector $\boldsymbol{x}^{(0)}$)

- Only method, which is applicable for large number of parameters (e.g. 10^6),
- no matrix $\boldsymbol{A}^{\#}$, no error propagation.

Methods

Correction factor method (CFM) and others: without any matrix and matrix operation, without use of data covariance matrix $\boldsymbol{V}_y$, requiring measurement y close to true value x, with the same bins for measurement and unfolded result. Often with intermediate smoothing/fit.

Landweber iteration: Matrix $\boldsymbol{A}$ unchanged, may be large and sparse:

$$\boldsymbol{x}^{(k)} := \boldsymbol{x}^{(k-1)} + \omega \boldsymbol{A}^{\mathrm{T}} \left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \right) \qquad k = 1, 2, \ldots \qquad 0 < \omega < 2 \left\| \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \right\|_2^{-1}$$

Matrix $\boldsymbol{A}$ remains unchanged, used only in product $\boldsymbol{A}\boldsymbol{x}^{(k-1)}$.

Iterative methods – how do they work? Example: Landweber iteration

$$\boldsymbol{x}^{(k)} := \boldsymbol{x}^{(k-1)} + \omega \boldsymbol{A}^{\mathrm{T}} \left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \right) \qquad k = 1, 2, \ldots \qquad 0 < \omega < 2 \left\| \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \right\|_2^{-1}$$

Surprisingly, solution similar to Thikhonov regularization, with an *implicit* filter factor $f_i^{(k)}$ in iteration k for the i-th eigen-contribution:

$$f_i^{(k)} = 1 - (1 - \omega\lambda_i)^k \qquad i = 1, 2, \ldots n$$

$$f_i^{(k)} \approx \begin{cases} k\,(\omega\lambda_i) & \text{for } \lambda_i \ll 1/\omega \\ 1 & \text{for large } \lambda_i \end{cases} \qquad \text{i.e. } f_i^{(k)} \ll 1 \quad \text{unless } k \text{ very large}$$

⇒ The iteration number k plays the role of a regularization parameter:

- convergence is fast for components with a large eigenvalue λ_j, and
- very slow for components with small eigenvalues: result after few iterations is still close to start vector $\boldsymbol{x}^{(0)}$.
- Without smoothing the result will be, after a large number of iterations, the unique (oscillating?) solution of the linear system ⇒ stop early(!), under which condition?
- Intermediate smoothing even removes the components with small eigenvalues!

Example of MC test in HEP publication: smoothing with 3rd degree polynomial ⇒ means only 4 (or 3) degress of freedom, for test distributions with linear and parabolic shapes (only 2 and 3 degrees of freedom), but 9 points are shown: strong positive correlation in result!

Comparison of unfolding methods

Property	Iterative methods e.g. CFM, Landweber	Regularized unfolding with $m > n$
meas. errors taken into account	no	yes
unbiased w.r.t model	no*)	yes
small bias w.r.t. general requirement	?	[yes]
matrix $\boldsymbol{A}^{\#}$, error propagation $\rightarrow \boldsymbol{V}_x$	no	yes
small bin-to-bin correlations	no	yes
simple	yes	no

*) unbiased for perfect model

- *"– for the estimate by CFM it is hard to calculate noise characteristics and bias in the estimate because of nonlinearity of the procedure"* (V.B.Anykeyev et al., NIM A 322 (1992) 280-285)
- *"Correction factors – a disaster. ... The data will tend to follow the MC that gave you the correction factors ..."* (Roger Barlow, SLUO Lecture 9 (2000) SLAC)

Summary

For every complex problem,
there is a solution that is simple, neat, and wrong.
H.L. Mencken

Experimental data, measured with finite resolution and limited acceptance, require unfolding to allow a correct interpretation:

- Statistical errors are increased, if there are migration effects.
- Number of bins of unfolded distribution is in general small, and has to be adjusted to the size of migration effects – otherwise correlations between bins will be large.
- Standard methods are based on
  - **Least squares (for Gaussian errors)**, or
  - **ML with Poisson distribution (for counting data, events)**, supplemented by
  - **regularization terms**,

  using orthogonalization methods (singular values or diagonalization), and allow to control a potential bias and to propagate the measurement errors.

Examples

Iterative methods are usually applied with intermediate "**smoothing**"

Example (1)

- *. . . requires the double differential cross section model used in the simulation to be sufficiently close to the data . . .*
- *. . . MC event reweighting procedure which converges after one iteration . . .*
- *. . . the measured double differential cross section is fitted by a new parametrization and the analysis of the MC events is repeated using an additional weight factor . . .*

Example (2)

- *. . . to produce stable results . . . smooth the results of the unfolding before feeding them in the next step as "initial probability" . . . the method is theoretically well grounded . . .*

Example (3)

- Four test distributions are tested by MC, two with straight lines (polynomial of 1^{st} degree) and two with a parabolic shape (polynomial of 2^{nd} degree) . . .
- *. . . a rough smoothing has been performed for all of them by a polynomial of 3^{rd} degree . . .*
- *. . . after a 20-step unfolding with intermediate smoothing . . . No oscillations are present and the results do not change with the increasing the number of steps indicating that the procedure has converged.*

Contents