# Unfolding

---

Volker Blobel – University of Hamburg

March 2005

## The unfolding problem

Reconstruct the distribution $f(x)$ of the true variable $x$ from the measured distribution $g(y)$ of the quantity $y$, which is related to the true variable $x$, based on the knowledge of the resolution (or migration) function $A(y, x)$.

$$g(y) = \int_{\Omega} A(y,x) f(x) \mathrm{d}x \qquad \text{or short} \qquad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \quad \text{or} \quad \boldsymbol{y} \cong \boldsymbol{A}\boldsymbol{x}$$

where

| | | | | | |
|---|---|---|---|---|---|
| $g(y)$ | = | measured distr. | $\boldsymbol{y}$ | = | measured histogram |
| $f(x)$ | = | ideal (true) distr. | $\boldsymbol{x}$ | = | true histogram |
| $A(y,x)$ | = | resolution fcn. | $\boldsymbol{A}$ | = | resolution/migration matrix |

The resolution matrix $\boldsymbol{A}$ has usually be determined from the $x/y$-pairs of a sample of MC events.

In matrix notation: determine vector $\boldsymbol{x}$ from the measured vector $\boldsymbol{y}$, with a known matrix $\boldsymbol{A}$ - i.e. solve a linear equation.

# Examples for a migration matrix

A Matrix $\boldsymbol{A}$, depending on a single parameter $\varepsilon$ (= migration parameter):

$$\boldsymbol{A} = \begin{pmatrix} 1-\varepsilon & \varepsilon & 0 & 0 & 0 \\ \varepsilon & 1-2\varepsilon & \varepsilon & 0 & 0 \\ 0 & \varepsilon & 1-2\varepsilon & \varepsilon & 0 \\ 0 & 0 & \varepsilon & 1-2\varepsilon & \varepsilon \\ 0 & 0 & 0 & \varepsilon & 1-\varepsilon \end{pmatrix}$$

Red elements show migration probability from second true bin into three bins of the measured distribution.

The matrix is symmetric and therefore has real eigenvalues.

# Solution of $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by inversion — First attempt

A direct solution is possible with inversion of the matrix $\boldsymbol{A}$:

$$\begin{array}{ll} \text{estimate} & \widehat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y} \\ \text{error propagation} & \boldsymbol{V}(\widehat{\boldsymbol{x}}) = \boldsymbol{A}^{-1}\boldsymbol{V}_y\left(\boldsymbol{A}^{-1}\right)^T \end{array}$$

The method has good statistical properties – no bias:

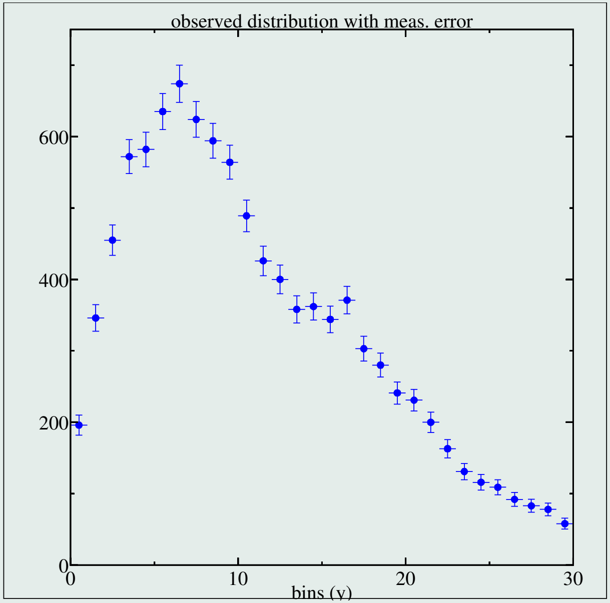
$$E[\boldsymbol{x}] = \boldsymbol{A}^{-1}E[\boldsymbol{y}] = \boldsymbol{A}^{-1}\boldsymbol{A}E[\boldsymbol{x}] = \boldsymbol{x}$$

In practice the result is satisfactory for a matrix $\boldsymbol{A}$ with dominating diagonal.

However: the results looks terrible if the matrix $\boldsymbol{A}$ describes a large migration to neighbour bins.

# An example for a measured histogram

…using migration parameter $\varepsilon = 0.24$, i.e. 52 % of true events remain in the same bin, and for 10 000 events.

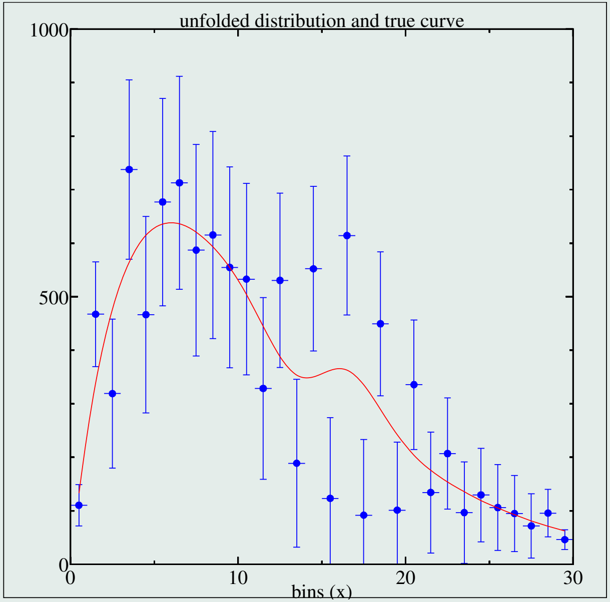


Note the small structure in the center:

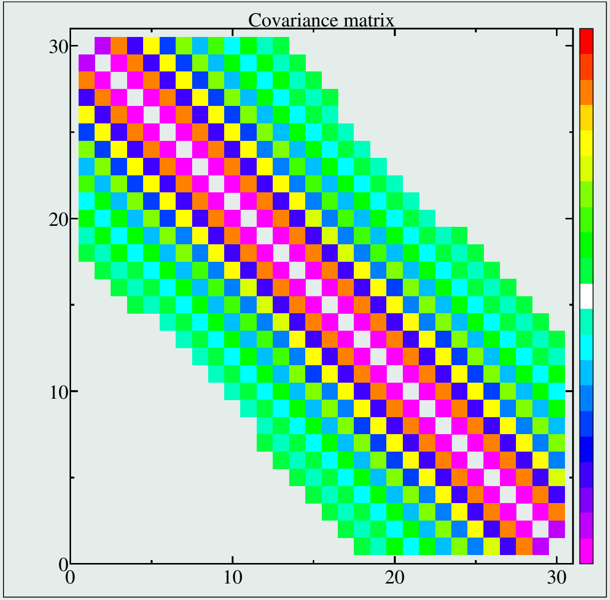
- It may be just a statistical fluctuation (→ smooth after unfolding)
- If it is a real structure in the distribution, then the true peak has to be higher!

# Result of solution by inversion

Reconstructed data points and true curve



Correlation coefficients $\rho ij$



Highly fluctuation data points due to large negative correlations, caused by limited resolution. Correlation coefficients $\rho_{ij}$ with $|\rho_{ij}| > 0.05$ are shown by colour boxes: here the coefficients $\rho_{i,i+1}$ between neighbour bins are $\approx -0.95$.

# Solution of $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by orthogonal decomposition

Decomposition of symmetric matrix $\boldsymbol{A}$: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ with diagonal matrix $\boldsymbol{D}$ of eigenvalues $\lambda$ and $\boldsymbol{U}^T\boldsymbol{U} = \mathbf{1}$. Matrix $\boldsymbol{U}$ contains eigenvector $\boldsymbol{u}_j$ in $j$-th column.

Transformation to new basis:

$$\begin{array}{lrcl} \boldsymbol{U}^T \cdot \;\; | & \boldsymbol{y} & \cong & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{x} \\ & \boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y} & \cong & \boldsymbol{D}\left(\boldsymbol{U}^T\boldsymbol{x}\right) = \boldsymbol{D}\boldsymbol{b} \qquad\qquad \boldsymbol{b} = \boldsymbol{D}^{-1}\boldsymbol{c} \end{array}$$

Vector $\boldsymbol{y}$ is transformed to vector $\boldsymbol{c}$ using matrix $U$ (eigenvectors $\boldsymbol{u}_j$):

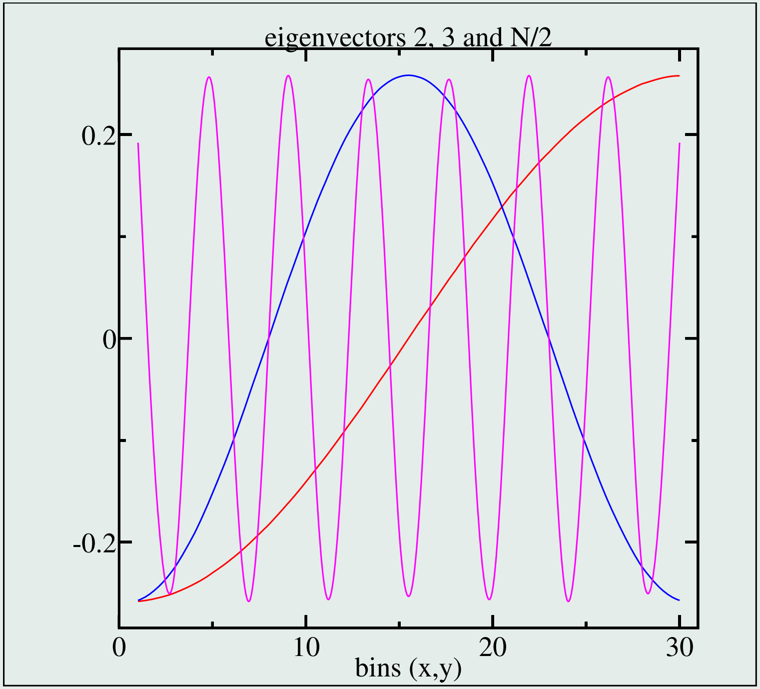
$$\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y} \qquad \text{or} \qquad c_j = \boldsymbol{u}_j^T\boldsymbol{y} \qquad j = 1 \ldots n$$

Unfolding is simply multiplication/division of coefficients by eigenvalues (and depends only an matrix $\boldsymbol{A}$):

$$\text{Unfolding} \qquad \boldsymbol{y} \rightarrow \boldsymbol{x} \qquad c_j \rightarrow \frac{c_j}{\lambda_j} = b_j = \boldsymbol{u}_j^T\boldsymbol{x} \qquad j = 1 \ldots n$$

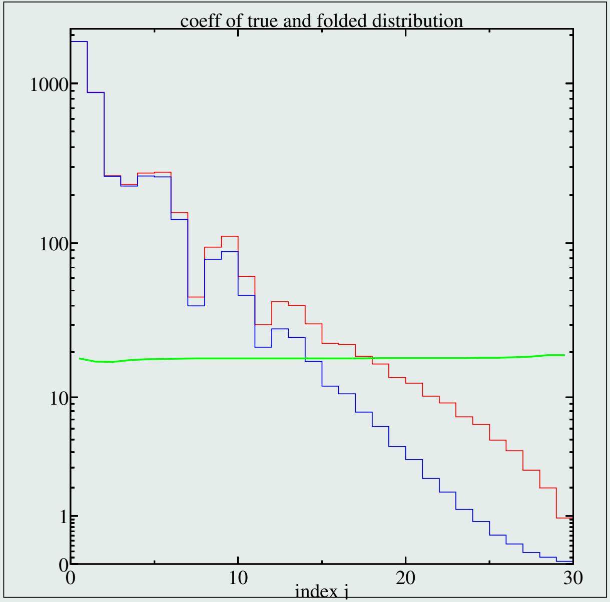
# Three eigenvectors in unfolding

Condition number of matrix is $\lambda_{\max}/\lambda_{\min} = 24$. The migration parameter is $\varepsilon = 0.24$, i.e. 52 % of the entries after folding remain in the correct bin.

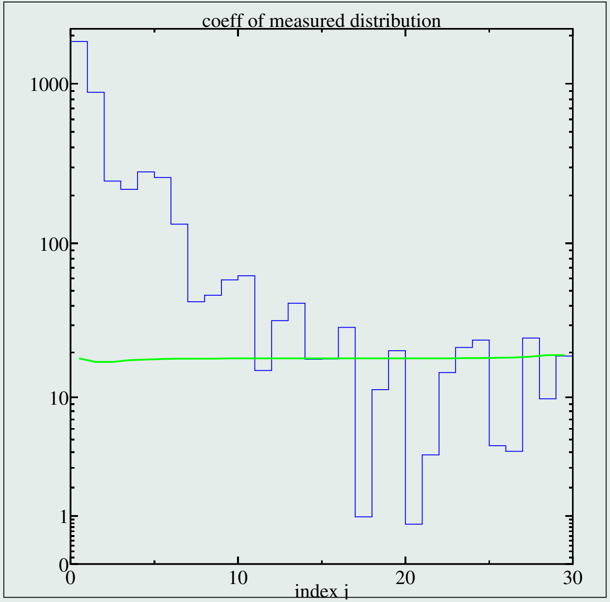


# Amplitudes … corresponding to eigenvectors

… for true distribution and folded distribution



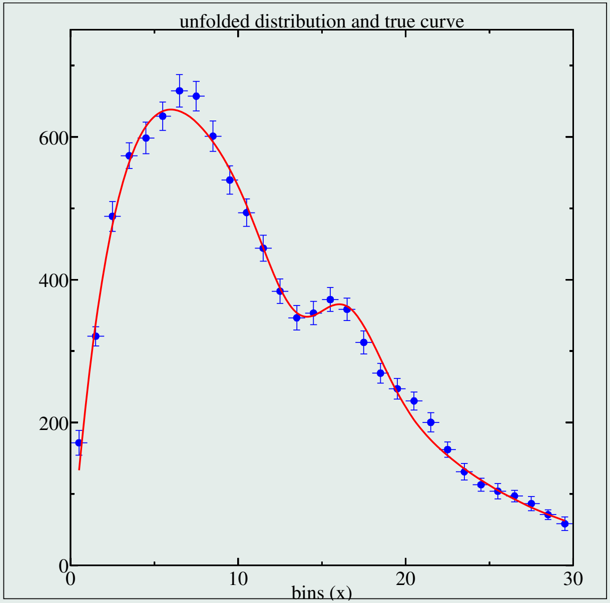
… from measurement

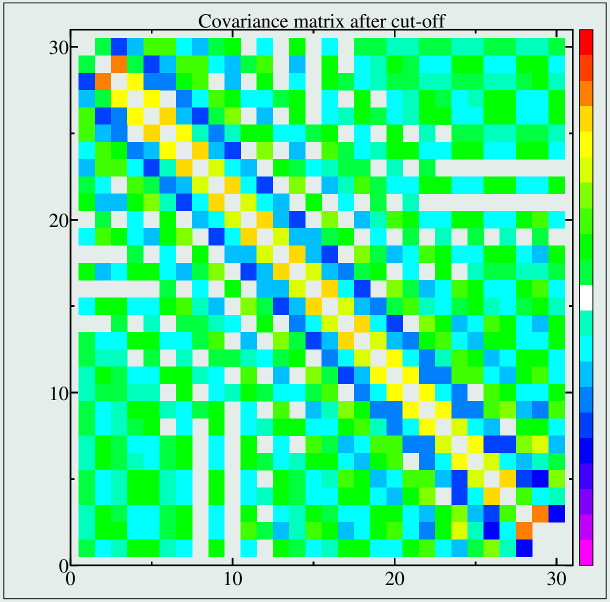


Folded amplitudes are measured and can be transformed to reconstruct the true amplitudes.
Green line represents statistical errors (noise level). True and folded amplitudes below the noise level can not be reconstructed.

# Solution with cut-off … at $N/2$ coefficients

Take only the significant first 15 amplitudes to reconstruct the distribution with 30 data points.

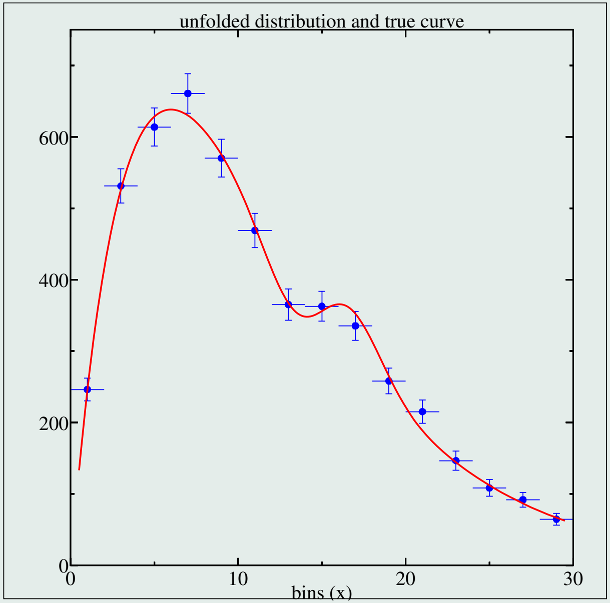


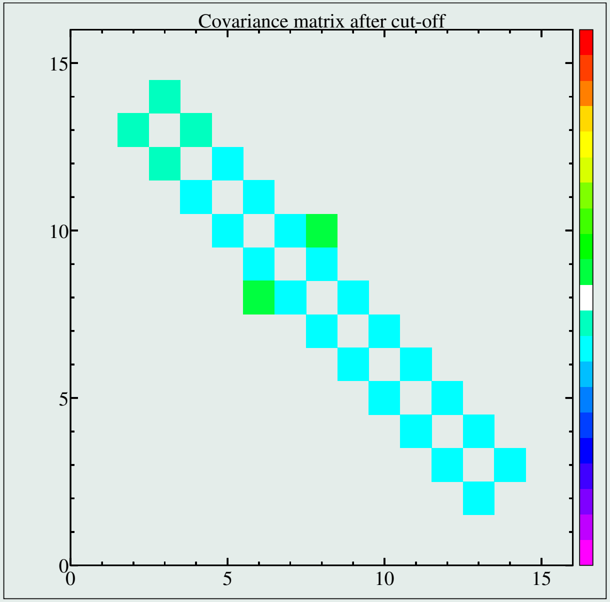


Covariance matrix has rank 15 and coefficients $\rho_{i,i+1}$ between neighbour bins are large and positive ($\approx +0.6$): “statistical” errors are smaller than original errors!

# Solution with $N/2$ data points

If two bins are combined to one, the distribution has a covariance matrix with full rank.





All correlation coefficients are small, even between neighbour bins ($|\rho_{i,i+1}| < 0.2$) – but at the cost of a reduced number of points.

The standard method in particle physics to correct for the limited resolution is explained *in words* (no mathematical formula):

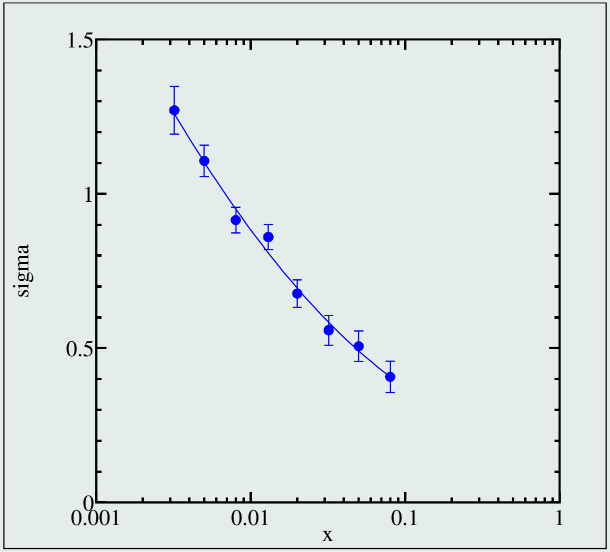
> . . . The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are large at large $x$ where the structure functions vary rapidly with $x$. We proceed by assuming a "true" structure function and calculate by Monte Carlo simulation, on the basis of the known experimental resolution functions, the result to be expected in the apparatus. **By iteration a "true" distribution which reproduces the experimental result is found.** The "unsmearing factor" is the ratio of Monte Carlo events for any particular $(x, Q^2)$ bin in the "true" distribution divided by those in the resolution smeared distribution. If this factor differs from unity by more than 30 %, the bin is not retained. . . . [**?**]

The method above is correct, if the "true" distribution is known without error.

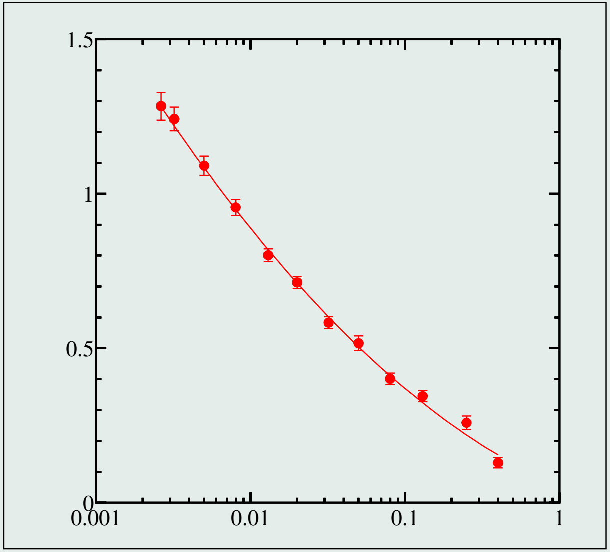
## Published data on deep inelastic scattering

Shown are the total (statistical and "uncorrelated" systematic) errors. In addition there are "correlated" systematic errors and a normalisation error of 1.8 % and 1.5 %, resp. The curve is a fitted parabola, with a $\chi^2$, that is better than expected (the data are rather smooth).

1998/1999 data (16.4 $\text{pb}^{-1}$)
at $Q^2 = 200$ GeV$^2$



1999/2000 data (65.2 $\text{pb}^{-1}$)
at $Q^2 = 200$ GeV$^2$



Unsmearing corrections are done based on earlier fits; bins are required to have stability and purity of $> 30\%$.

# Unfolding