Unfolding with data fitting

Nikolai D. Gagunashvili

University of Akureyri, Iceland

Unfolding framework meeting, July 2011

Analysis Center, Helmholtz Alliance "Physics at the Terascale"

The probability density function (PDF) $P(x')$ of a reconstructed characteristic x' of an event obtained from a detector with finite resolution and limited acceptance can be represented as

$$P(x') = \int_{\Omega} p(x)A(x)R(x,x')\,dx, \tag{1}$$

where $p(x)$ is the true PDF, $A(x)$ is the acceptance of the setup, i.e. the probability of recording an event with a characteristic x, and $R(x,x')$ is the experimental resolution, i.e. the probability of obtaining x' instead of x after the reconstruction of the event. The integration in (1) is carried out over the domain Ω of the variable x.

To solve unfolding problem or find $p(x)$ from (1). Let us represent $p(x)$ as:

$$p(x) = a_0 + \sum_{i=1}^{i=m} a_i K(\frac{x - x_i}{\lambda}) \tag{2}$$

where a_i are positive parameters, $K(\frac{x-x_i}{\lambda})$ are kernel with position of center of kernel x_i and scale parameter λ.

Kernels are widely used in non-parametric regression analysis.

Examples of kernels:

$\frac{3}{4\lambda}(1 - \frac{(x-x_i)^2}{\lambda^2})1_{\{|x-x_i|\leq\lambda\}}$ – Epanechnikov Kernel;

$\frac{35}{32\lambda}(1 - \frac{(x-x_i)^2}{\lambda^2})^3 1_{\{|x-x_i|\leq\lambda\}}$ – Triweight Kernel;

$\frac{1}{\lambda\pi}(\frac{\lambda^2}{\lambda^2+(x-x_i)^2})$ – Cauchy Kernel;

$\frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2\lambda^2}}$ – Gaussian Kernel.

Let us substitute (2) to basic equation (1):

$$P(x') = a_0 \int_\Omega A(x)R(x,x')\,dx + \sum_{i=1}^{i=m} a_i \int_\Omega K(\frac{x-x_i}{\lambda})A(x)R(x,x')\,dx.$$

Elements of equation
$\int_\Omega A(x)R(x,x')\,dx$, $\int_\Omega K(\frac{x-x_i}{\lambda})A(x)R(x,x')\,dx$ can be calculated in advance using Monte-Carlo.
After that we have liner fit problem and find estimators $\hat{a}_i$ and unfolded distribution:

$$p(x) = \hat{a}_0 + \sum_{i=1}^{i=m} \hat{a}_i K(\frac{x-x_i}{\lambda}) \qquad (3)$$

The method described above is now illustrated with an example proposed by Blobel. We take a true distribution

$$\phi(x) \propto \sum_{i=1}^{3} A_i \frac{C_i^2}{(x - B_i)^2 + C_i^2} \tag{4}$$

with the same parameters as in a previous study, where x is defined on the interval $[0, 2]$.

A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3
1	10	5	0.4	0.8	1.5	2	0.2	0.2

An experimentally measured distribution is defined as

$$P(x) = \int_0^2 p(x')A(x')R(x,x')dx' \tag{5}$$

where the acceptance function $A(x)$ is

$$A(x) = 1 - \frac{(x-1)^2}{2} \tag{6}$$

and

$$R(x,x') = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-x'+0.05x'^2)^2}{2\sigma^2}) \tag{7}$$

is the detector resolution function with $\sigma = 0.1$.

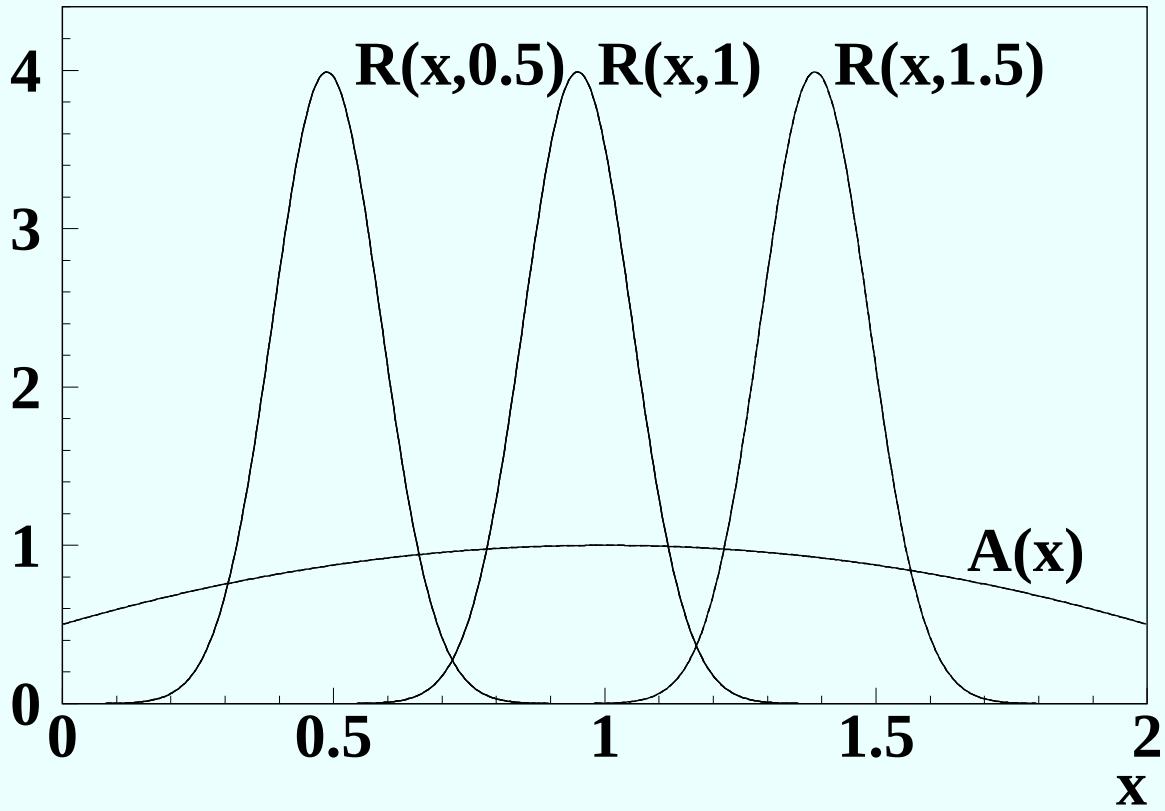


Figure: The acceptance function $A(x)$ and resolution function $R(x, x')$ for $x' = 0.5, 1.0$ and 1.5.

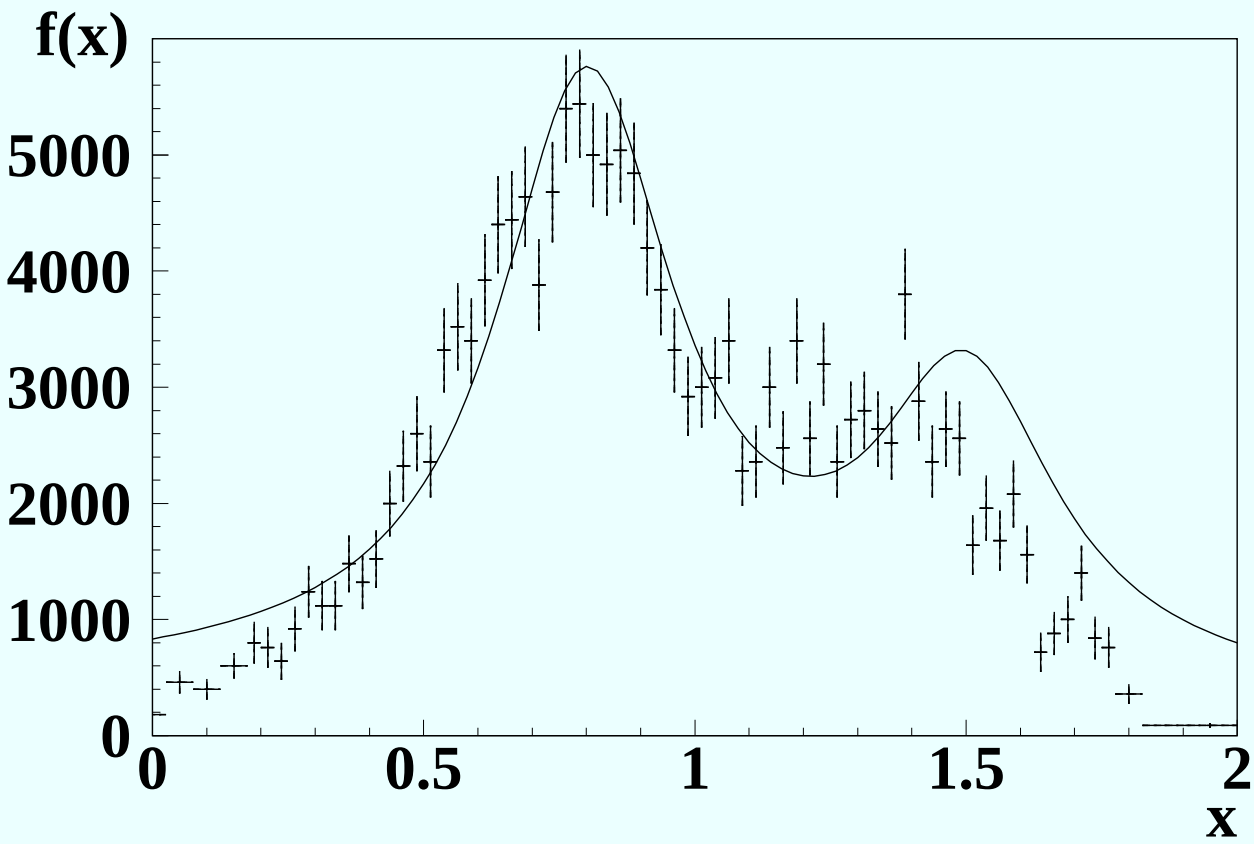


Figure: The measured distribution $P(x)$ (number of events divided on bin size). The true distribution $p(x)$ is shown as curve.

λ	χ^2	n
.25	52.0	11
.30	55.1	9
.35	57.5	9
.40	57.0	7
.45	56.6	7
.50	56.8	7
.55	68.3	4
.60	115.	4

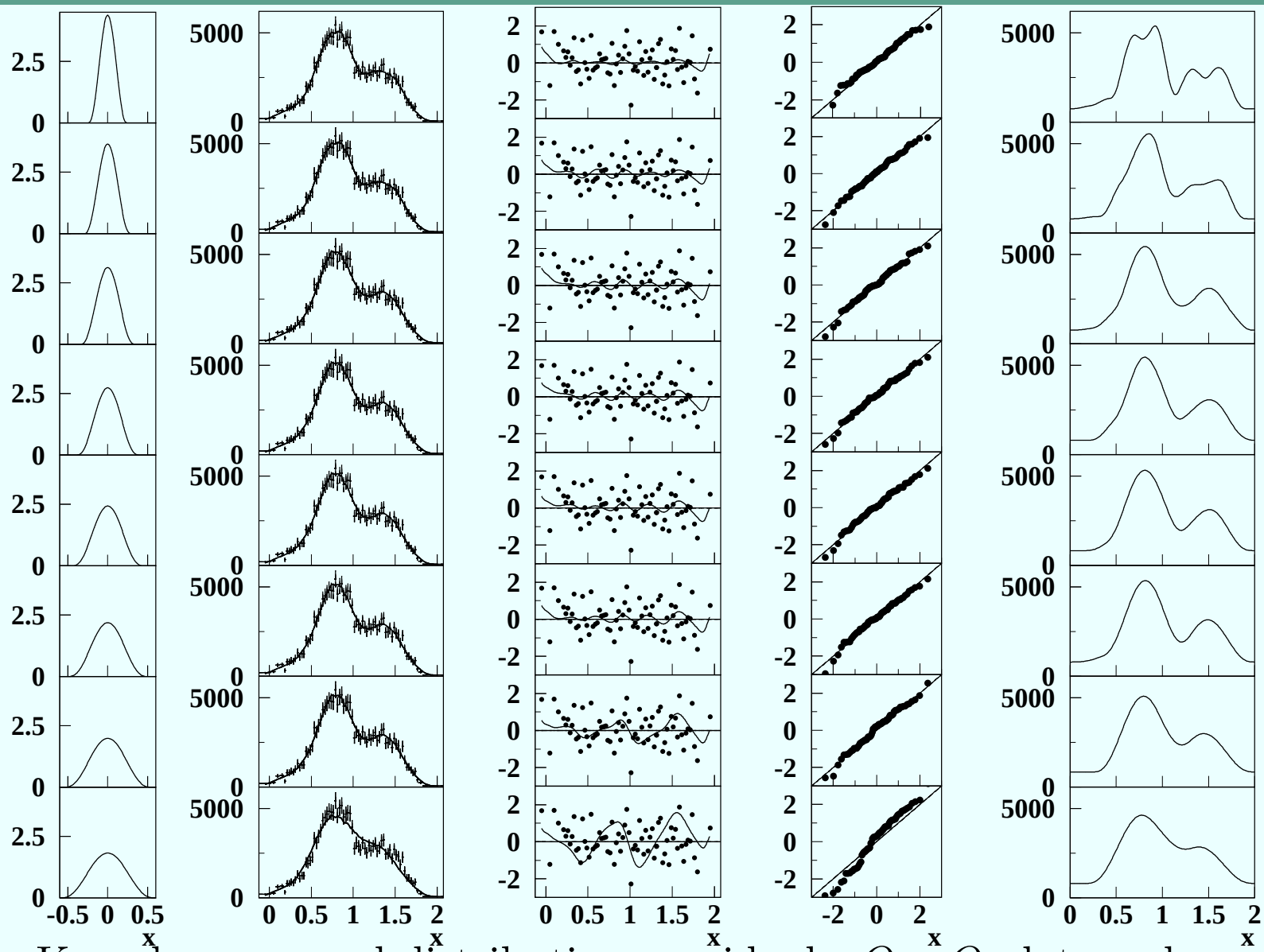


Figure: Kernels, measured distributions, residuals, $Q - Q$ plots and unfolded distributions for different bandwidth λ of kernels.

λ	χ^2	n
.25	52.0	11
.30	55.1	9
.35	57.5	9
.40	57.0	7
.45	56.6	7
.50	56.8	7
.55	68.3	4
.60	115.	4

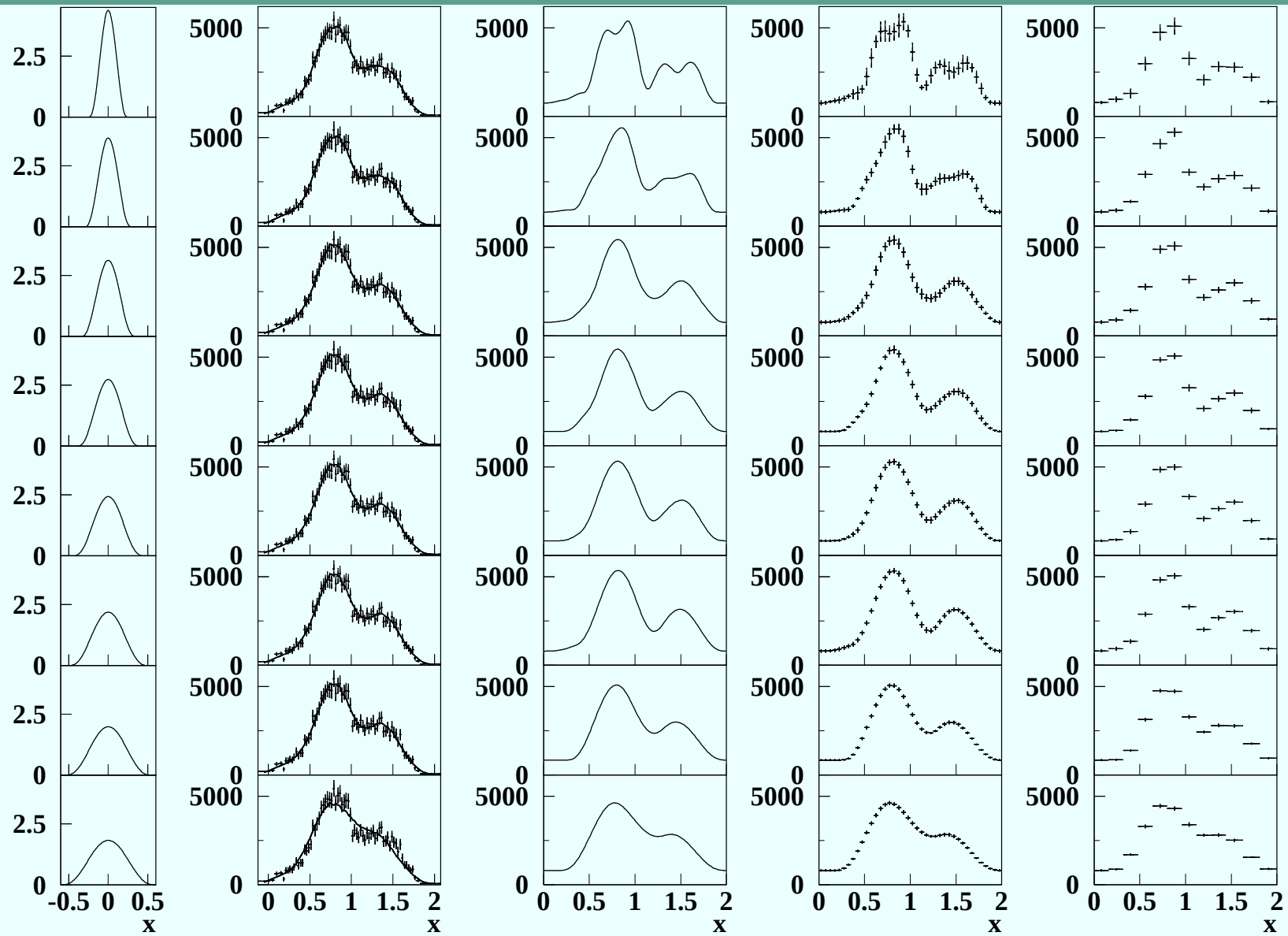


Figure: Kernels, measured distributions, unfolded distributions, unfolded distributions(40 bins), unfolded distributions (12 bins).

To investigate the statistical properties of the unfolding procedure, 1000 simulation runs were performed for the same true distribution. The unfolded distribution was calculated for each measured distribution. The following quantities were calculated:

- Exact value of the components of the true distribution $p_i = 5000 \int_{x_i}^{x_{i+1}} p(x)dx/(x_{i+1} - x_i)$ where x_{i+1} and x_i are the bounds of ith bin.
- Average value of the unfolded distribution $\bar{\hat{p}}_i = \sum_{j=1}^{1000} \hat{p}_i(j)/1000$, where j is the run number.
- Bias for components of the unfolded distribution $B\hat{p}_i = \bar{\hat{p}}_i - p_i$
- Standard deviation s_i for the unfolded distribution components $s_i = \sqrt{\sum_{j=1}^{1000} (\hat{p}_i(j) - \bar{\hat{p}}_i^2)/1000}$.
- Mean estimated error $\hat{\delta}_{ii}$ for the unfolded distribution components $\bar{\hat{\delta}}_{ii} = \sum_{j=1}^{1000} \hat{\delta}_{ii}(j)/1000$.
- Bias for errors in the unfolded distribution components $B\hat{\delta}_{ii} = s_i - \bar{\hat{\delta}}_{ii}$.

- Mean Square Error: $MSE_i = \sum_{j=1}^{1000}(\hat{p}_i(j) - p_i)^2/1000$, it is known that $MSE_i = s_i^2 + (Bp_i)^2$
- Total Mean Square Error: $TMSE = \sum_i MSE_i = \sum_i s_i^2 + \sum_i (Bp_i)^2$

$\sum_i s_i^2$ – Total Variance (TV)

$\sum_i (Bp_i)^2$ – Total Squared Bias (TSB)

$TMSE = TV + TSB$

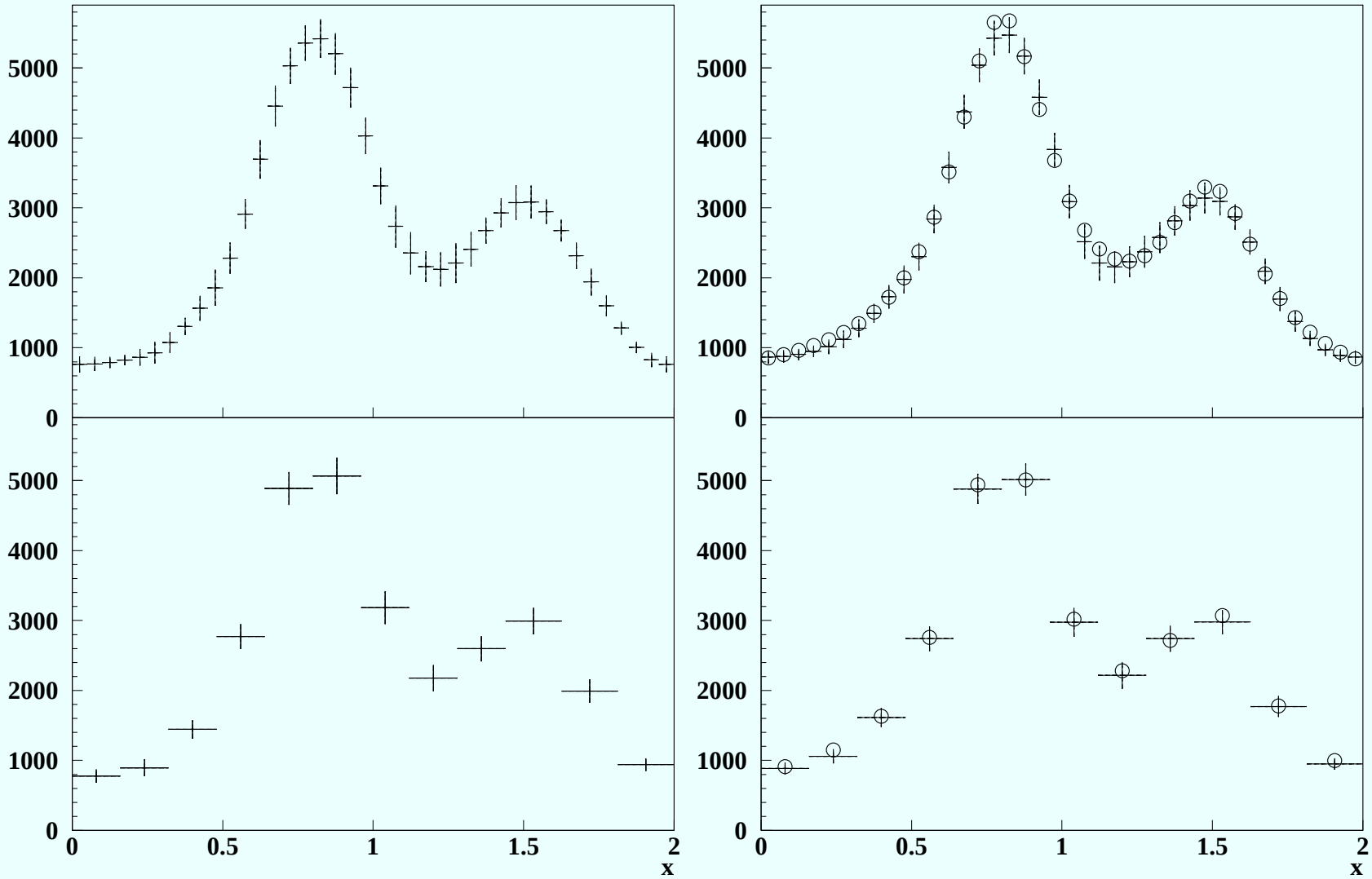
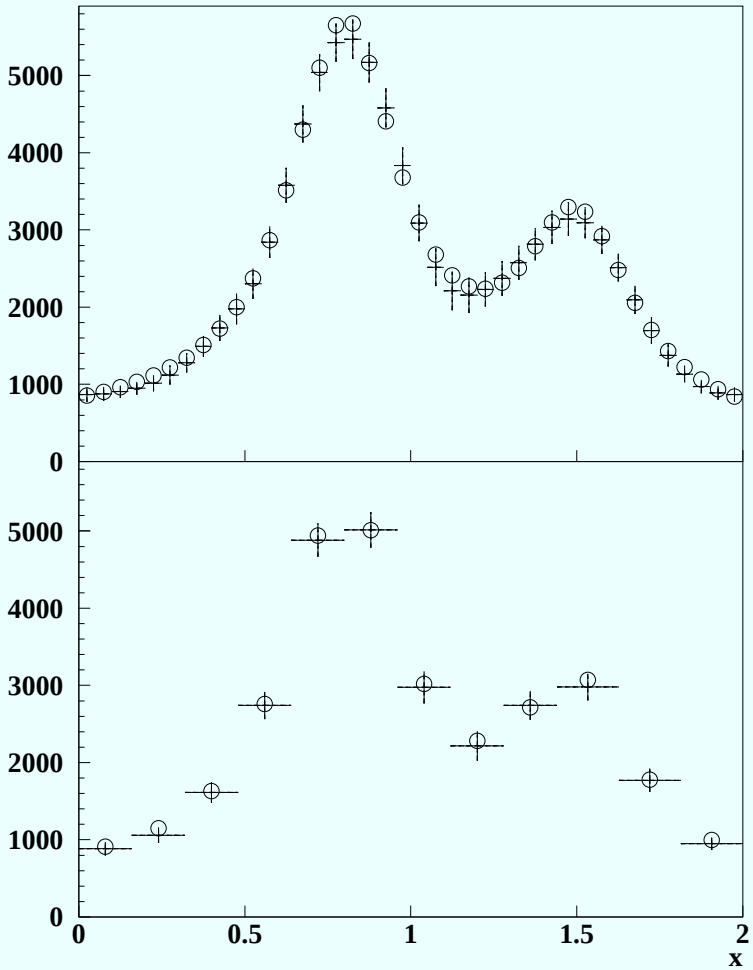
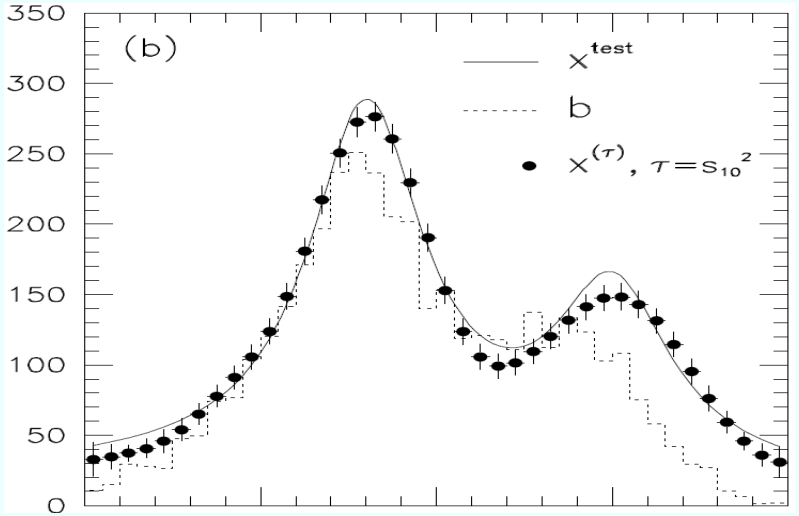


Figure: Unfolded distributions ($\lambda = 0.35$) and average unfolded distribution with average errors (1000 runs), the circle centers ($\odot$)





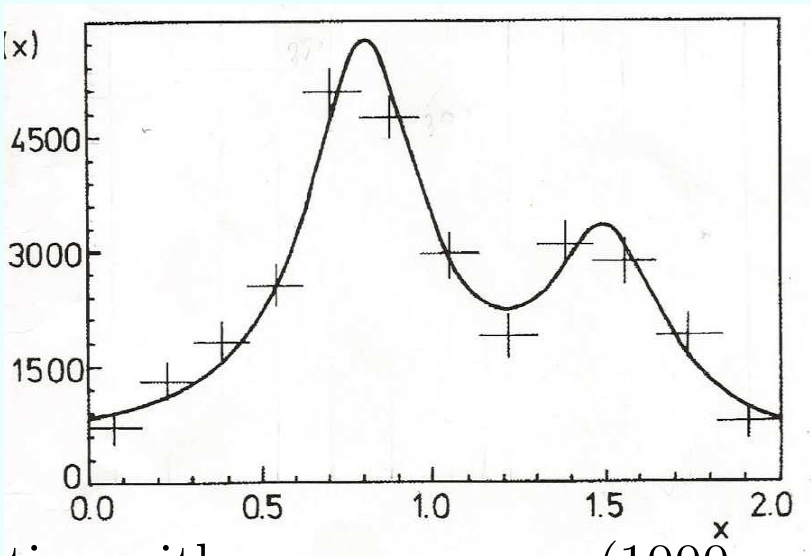


Figure: Average unfolded distribution with average errors (1000 runs) and unfolding results from A. Höcker, V. Kartvelishvili, NIM A372, 1996 and V. Blobel, CERN 85-02, 1985

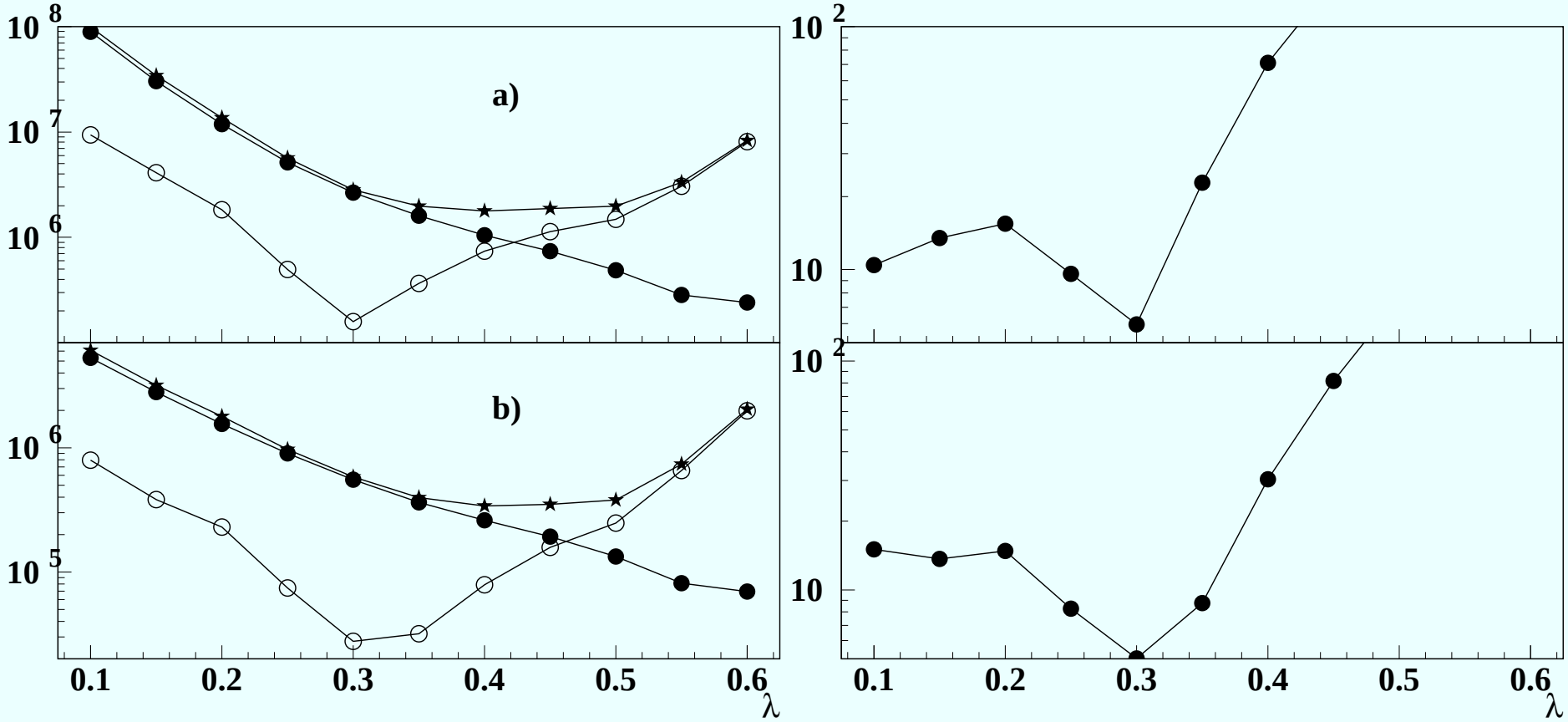


Figure: Total Mean Square Error ($\star$), total variance ($\bullet$), total squared bias ($\circ$) and ratio of total squared bias to total variance (%) for 40 bins (a) and for 12 bins (b)