Densities mixture unfolding for data obtained from detectors with finite resolution and limited acceptance

Nikolai Gagunashvili[1 2] Michael Schmelling[2]

[1]University of Akureyri, Akureyri, Iceland

[2]Max-Planck-Institut für Kernphysik, Heidelberg, Germany

ACAT, Prague, September 2014

vspace 0.5cm

Introduction

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

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

where

- $p(x)$ is the true PDF,
- $A(x)$ is the probability of recording an event with a characteristic x (the acceptance),
- $R(x'|x)$ is the probability of obtaining x' instead of x after the reconstruction of the event (the experimental resolution),

vspace 0.5cm

Solving main equation (1) or *unfolding* true distribution $p(x)$ is an underspecified problem and every approach to solve it requires a priori information about the solution. Different methods differ, directly or indirectly, in the use of this a priori information.

vspace 0.5cm

Description of the unfolding method

To solve the unfolding problem (1) a representation of the true distribution has to be chosen. It must be rather flexible to introduce a priori information.
Here we use the Mixture Density Model (MDM) for the true distribution $p(x)$:

$$p(x) = \sum_{i=1}^{s} w_i \, K_i(x, a_{1i}, ..., a_{li}), \qquad (2)$$

where

- $K_i(x, a_{1i}, ..., a_{li})$, $i = 1, \ldots s$, is the ith PDF with parameters $a_{1i}, ..., a_{li}$
- w_i is the fraction of the ith Probability Density Function in Mixture (PDFM).

vspace 0.5cm

Example of MDM is Gaussian Mixture Model (GMM) with PDFMs:

$$K_i(x, x_i, \lambda_i) = \frac{1}{\lambda_i \sqrt{2\pi}} \exp\left(-\frac{(x - x_i)^2}{2\lambda_i^2}\right) \tag{3}$$

that has two parameters:

- x_i the mean values (position)
- λ_i the standard deviation (width).

It is a rather flexible model for an approximation of a wide class of statistical distributions. The standard deviation λ_i of the PDFMs acts as a regularization parameter which allows for adjusting the smoothness of the result.

vspace 0.5cm

Substituting $p(x)$ represented by Eq. (2) into the basic Eq. (1) yields

$$P(x') = \sum_{i=1}^{s} w_i \int_{\Omega} K_i(x, x_i, \lambda_i) A(x) R(x'|x)\, dx. \qquad (4)$$

Weights w_i, positions x_i and standard deviations λ_i will be determined by the unfolding procedure.

vspace 0.5cm

In discrete form Eq. (4) is represented as:

$$\boldsymbol{P} = \mathbf{Q}\boldsymbol{w} + \boldsymbol{\epsilon}\,, \tag{5}$$

where

- $\boldsymbol{P}$ is the n-component vector of the experimentally measured histogram,
- $\boldsymbol{w} = (w_1, w_2, ..., w_s)^t$ is the s-component vector of weights
- $\mathbf{Q}$ is an $n \times s$ matrix with elements

$$Q_{ji} = \int_{c_{j-1}}^{c_j} K_i(x, x_i, \lambda_i)\, A(x)\, R(x'|x) dx \tag{6}$$

- $\boldsymbol{\epsilon}$ is an n-component vector of random residuals with expectation value $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and covariance matrix $\mathbf{C}$ with diagonal elements $\mathrm{Var}[\boldsymbol{\epsilon}] = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$, where σ_j is the statistical error of the measured distribution for the jth bin.

vspace 0.5cm

For a given set of PDFMs the weights $\boldsymbol{w}$ in Eq. (5) is determined such that it minimizes

$$X^2 = (\boldsymbol{P} - \mathbf{Q}\boldsymbol{w})^t \mathbf{C}^{-1} (\boldsymbol{P} - \mathbf{Q}\boldsymbol{w}) = \sum_{i=1}^{n} (P_i - \sum_{j=1}^{s} Q_{ij} w_j)^2 / \sigma_i^2 \quad (7)$$

with constrains

$$w_i \geq 0 \quad i = 1, ..., s. \quad (8)$$

The nonnegative least square problem is solved by picking the subset of w_i satisfying (8) such that (7) is smallest.

Numerical algorithm for solving this minimization problem developed by C. L. Lawson and R. J. Hanson. Subset of indices of components of estimation $\hat{\boldsymbol{w}}$ equal to 0 are calculated iteratively. The rest of indexes are defined subset of positive components of $\hat{\boldsymbol{w}}$.

Finally the positive components of solution is found by simple linear regression on the unconstrained subset of variables.

$$\hat{\boldsymbol{w}} = (\boldsymbol{\mathcal{Q}}^{\boldsymbol{t}}\mathbf{C}^{-\mathbf{1}}\boldsymbol{\mathcal{Q}})^{-1}\,(\boldsymbol{\mathcal{Q}}^{t}\mathbf{C}^{-\mathbf{1}})\,\boldsymbol{P}, \tag{9}$$

where $\boldsymbol{\mathcal{Q}}$ is submatrix of matrix $\mathbf{Q}$ corresponds to subset of indexes of positive components of solution.
When the method stops an estimate $\hat{p}(x)$ has been found, defined by subset of parameters $x_i, \lambda_i,\ i = 1, \ldots, k$ which are summed with non-zero weights $w_i, i = 1, \ldots, k$ to yield

$$\hat{p}(x) = \sum_{i=1}^{k} \hat{w}_i K_i(x, x_i, \lambda_i)\ . \tag{10}$$

vspace 0.5cm

The Prediction Error (PE) is defined as the average error in predicting new experimentally measured histogram $\boldsymbol{P}^{new}$ using predictor $\boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}$:

$$PE(\boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}) = E[\frac{1}{n}(\boldsymbol{P}^{new} - \boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}})^t \mathbf{C}^{-1}(\boldsymbol{P}^{new} - \boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}})], \qquad (11)$$

where expectation is over $\boldsymbol{P}^{new}$. Further predictor $\boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}$ we also will denote as $\hat{\boldsymbol{P}}$.

vspace 0.5cm

V-fold cross-validation error can be used to estimate $PE(\boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}})$. The set of data $\mathcal{U}$ are split into V subsets $\mathcal{U}_1, ..., \mathcal{U}_V$ with equal number of events. Let $\mathcal{U}^{(v)} = \mathcal{U} - \mathcal{U}_v$. Using minimization procedure and the data subset $\mathcal{U}^{(v)}$, form the predictors $\boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}^{(v)}$. The Cross Validation error (CV) is equal to

$$CV = \frac{1}{n}\sum_{v=1}^{V}(\boldsymbol{P}_v - \boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}^{(v)})^t \mathbf{C}^{-1}(\boldsymbol{P}_v - \boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}^{(v)}) \qquad (12)$$

where $\boldsymbol{P}_v$ is vector of histogram content for the subset of data $\mathcal{U}_v$. The cross-validation error is estimation of Prediction Error

$$CV = \widehat{PE}(\boldsymbol{\mathcal{Q}}\hat{\boldsymbol{w}}). \qquad (13)$$

Unfolding procedure has four steps:

vspace 0.5cm

First step

The positions $\{x_i\}$ of the PDFMs is taken randomly with uniform distribution on the range of x and with number of PDFMs that provide average distance between PDFM essentially lower than variance of PDFMs. Variance for all PDFMs is taken constant $\lambda_i = \lambda$ and we choose value of variance $\hat{\lambda}$ with minimal value of cross validation error

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, CV(\lambda) \tag{14}$$

.

vspace 0.5cm

Second step

The positions $\{x_i\}$ of the PDFMs is taken randomly with unfolded density obtained on the first step. Variances of PDFMs are taken different with $\{\hat{\lambda}_i\}$ as:

$$\hat{\lambda}_i = \frac{\hat{\lambda}}{\sqrt{\hat{p}(x_i)}} \tag{15}$$

where

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\ CV\left(\frac{\lambda}{\sqrt{\hat{p}(x_i)}}\right). \tag{16}$$

Here we use results of I. S. Abramson obtained for kernel estimations of PDFs. It was shown that by this way the value of bias of the kernel estimation can be decreased as well as in our case will be decreased bias of unfolded distribution.

Third step

Find better subset of PDFMs using non-negative garrote method by L. Brieman. Let $\{\hat{w}_j\}$ is the set of non-zero weights obtained on the second step and find $\{c_j\}$ that minimize

$$\sum_{i=1}^{n}(P_i - \sum_{j=1}^{s} \mathcal{Q}_{ij} c_j \hat{w}_j)^2/\sigma_i^2 \tag{17}$$

under the constraints

$$c_j \geq 0, \ \sum_{j=1}^{s} c_j \leq r \tag{18}$$

The $\tilde{w}_j(r) = c_j \hat{w}_j$ are the new estimation of weights. The garrote eliminate some of weights and shrinks other. Cross-validation is used for the choice optimal garrote parameter r.

vspace 0.5cm

Fourth step (Evaluation)

Asses the quality of fit with common tools used in regression analysis:

1. p-value of fit
2. analysis of the normalized residuals of the data
   1. as a function of the estimated value $\hat{\boldsymbol{P}}$
   2. as a function of the observed value x'
3. Q-Q plot: quantile of normalized residuals versus the theoretical quantile expected from a standard normal $\mathcal{N}(0,1)$ distribution.
4. Estimate statistical variations of unfolded distribution by bootstrap method.

vspace 0.5cm

Example 1

The method described above is now illustrated using an example proposed by V. Blobel. The true distribution, defined on the range $x \in [0, 2]$ is described by a sum of three Breit-Wigner functions

$$p(x) \propto \frac{4}{(x-0.4)^2+4} + \frac{0.4}{(x-0.8)^2+0.04} + \frac{0.2}{(x-1.5)^2+0.04} \tag{19}$$

from which the experimentally measured distribution is obtained by

$$P(x') \propto \int_0^2 p(x)A(x)R(x'|x)dx, \tag{20}$$

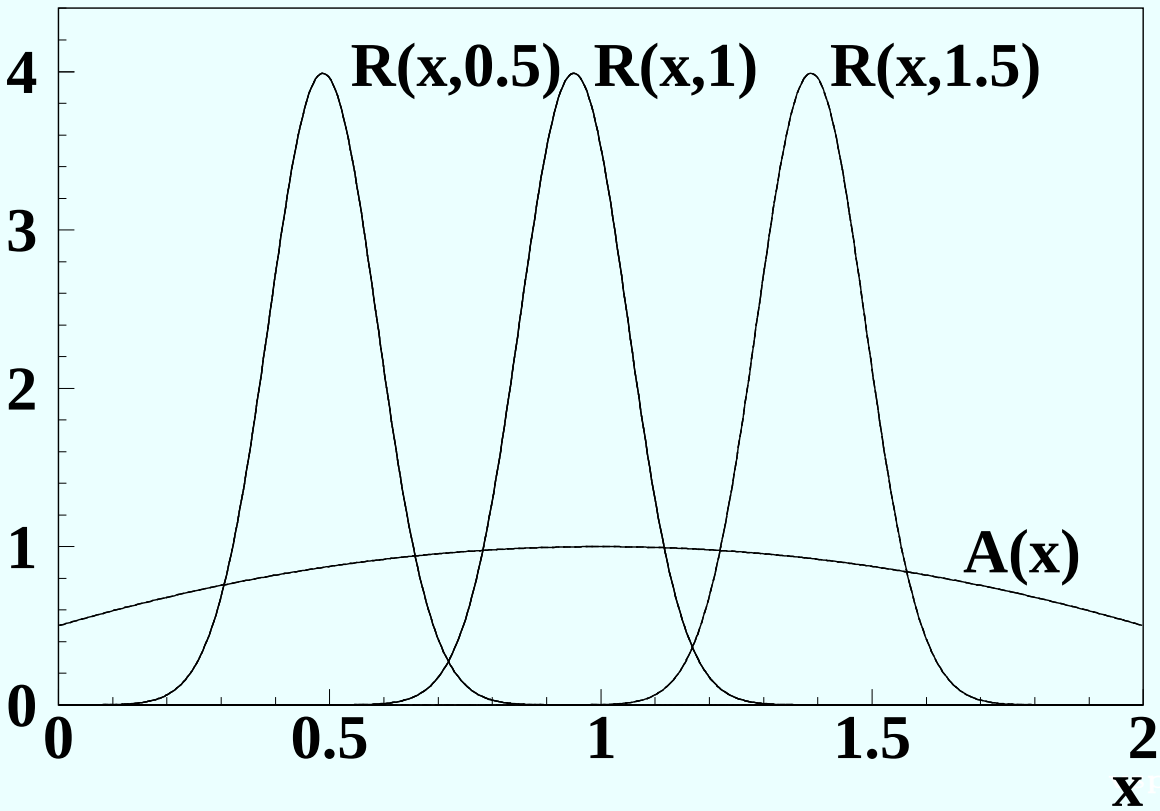
with an acceptance function $A(x)$

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

vspace 0.5cm

and a resolution function describing a biased measurement with gaussian smearing

$$R(x'|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x' - x + 0.05x^2)^2}{2\sigma^2}\right), \sigma = 0.1\ . \quad (21)$$



The measured distribution obtained by simulating a sample of $N = 5000$ events.
PDFMs was defined according B. W. Silverman for the PDF defined on the restricted interval in the form:

$$K_i(x, x_i, \lambda_i) \propto [\frac{1}{\lambda_i\sqrt{2\pi}} \exp\left(-\frac{(x-x_i)^2}{2\lambda_i^2}\right) + \frac{1}{\lambda_i\sqrt{2\pi}} \exp\left(-\frac{(x+x_i)^2}{2\lambda_i^2}\right) + \frac{1}{\lambda_i\sqrt{2\pi}} \exp\left(-\frac{(x-4+x_i)^2}{2\lambda_i^2}\right)] I_{\{x\in[0;2]\}}$$

where $I_{\{x\in[0;2]\}}$ the indicator function.

vspace 0.5cm

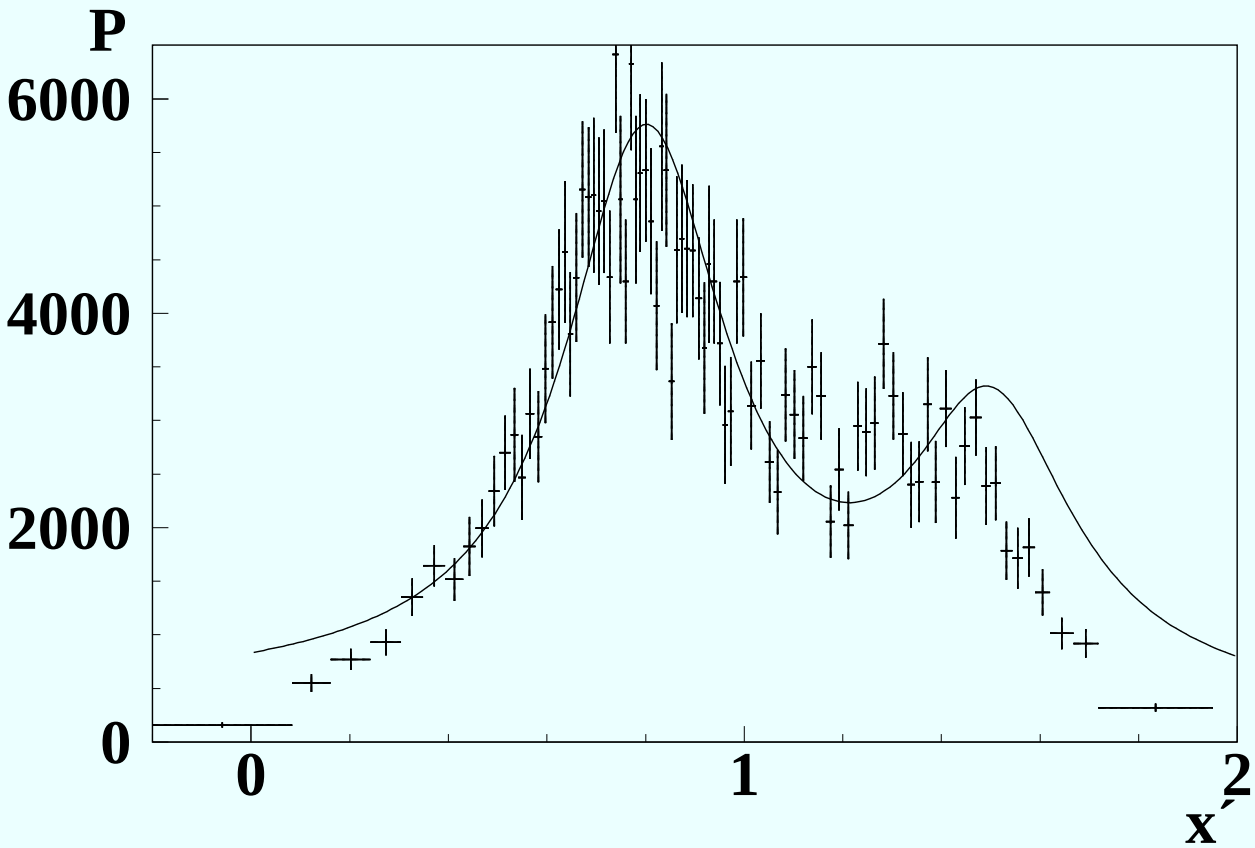


Figure: The measured distribution $P(x)$ (number of events divided on bin size). The true distribution $p(x)$ is shown as curve.

vspace 0.5cm

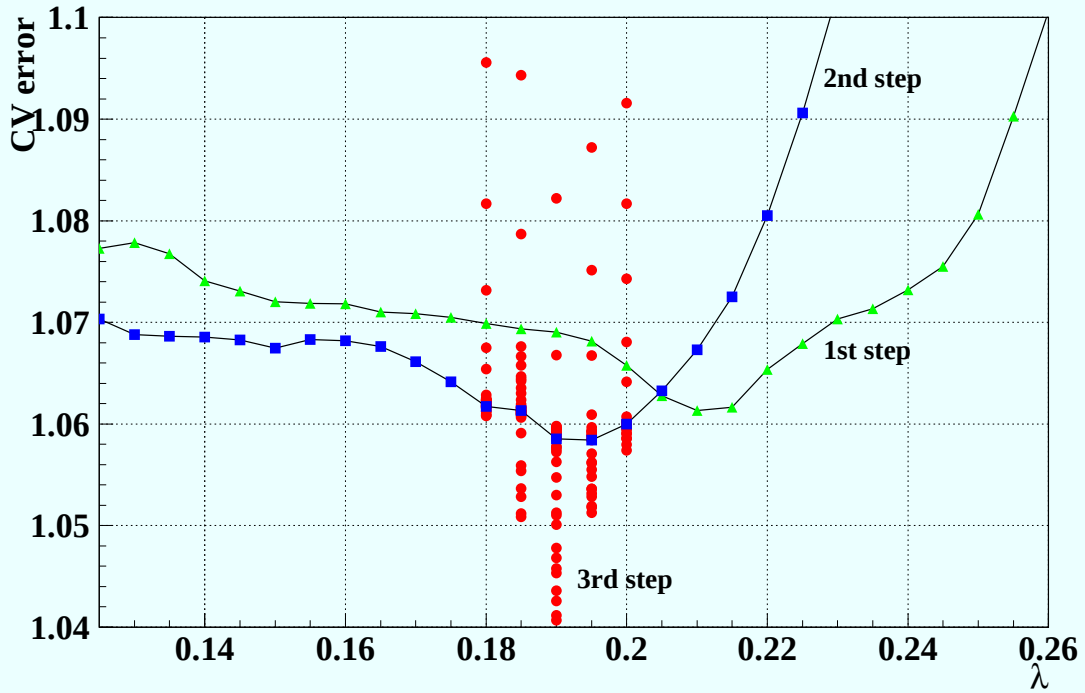


Figure: Cross Validation error for different values of λ

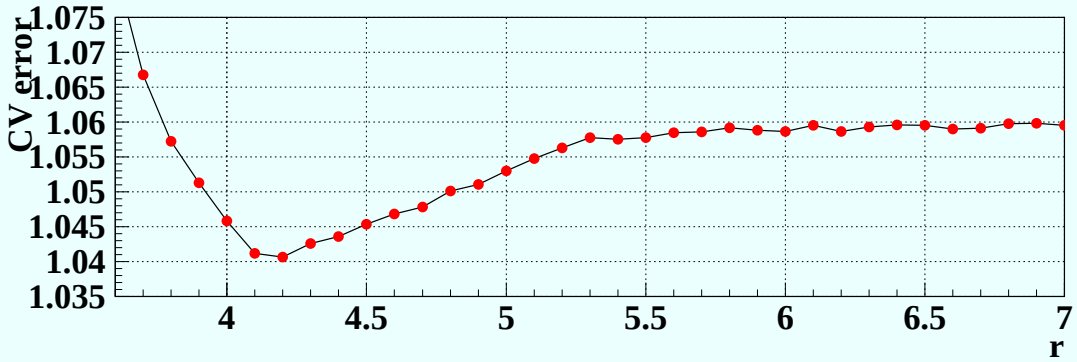


Figure: Cross Validation error for different values of r, $\lambda = 0.19$

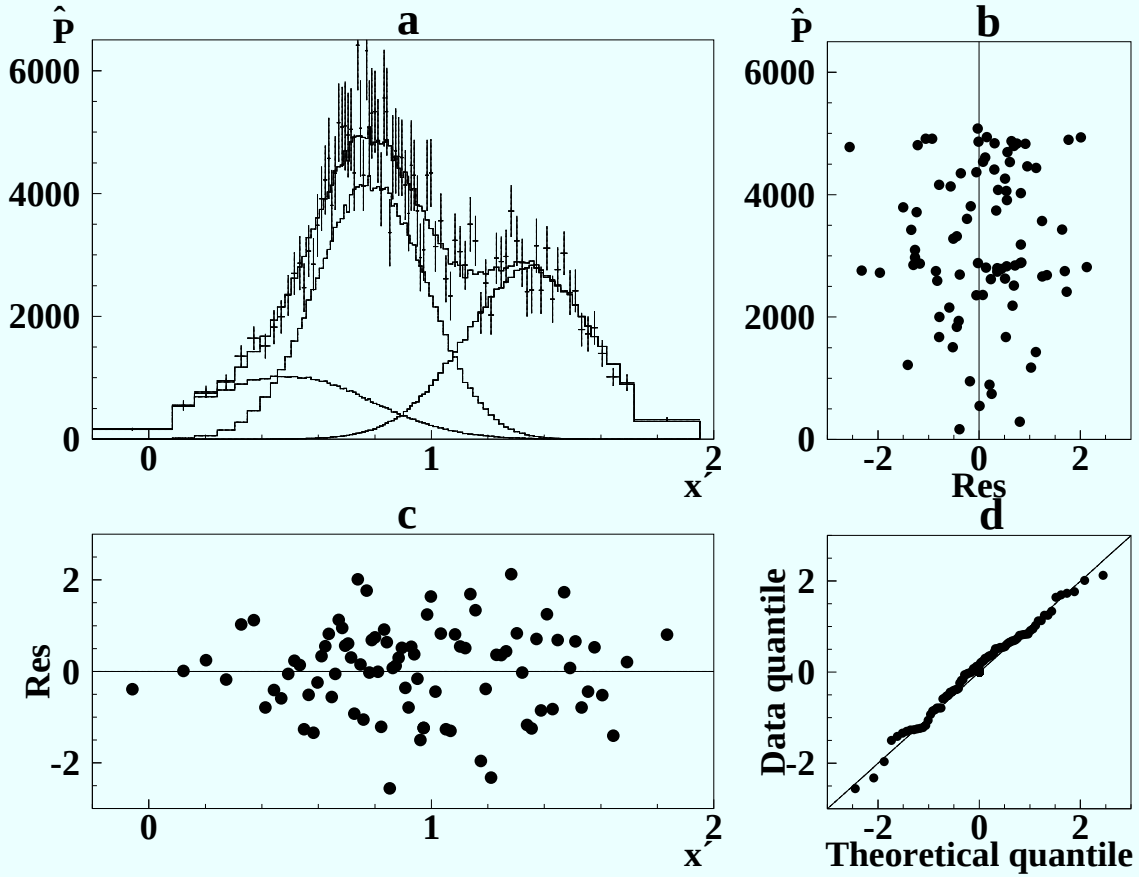


Figure: Illustration of the quality of the unfolding result. (a) folded PDFMs of the estimate of the true distribution compared to the measured distribution; (b) normalized residuals of the fit as a function of $\hat{\boldsymbol{P}}$; (c) normalized residuals as a function of x';
(d) quantile-quantile plot for the normalized residuals.

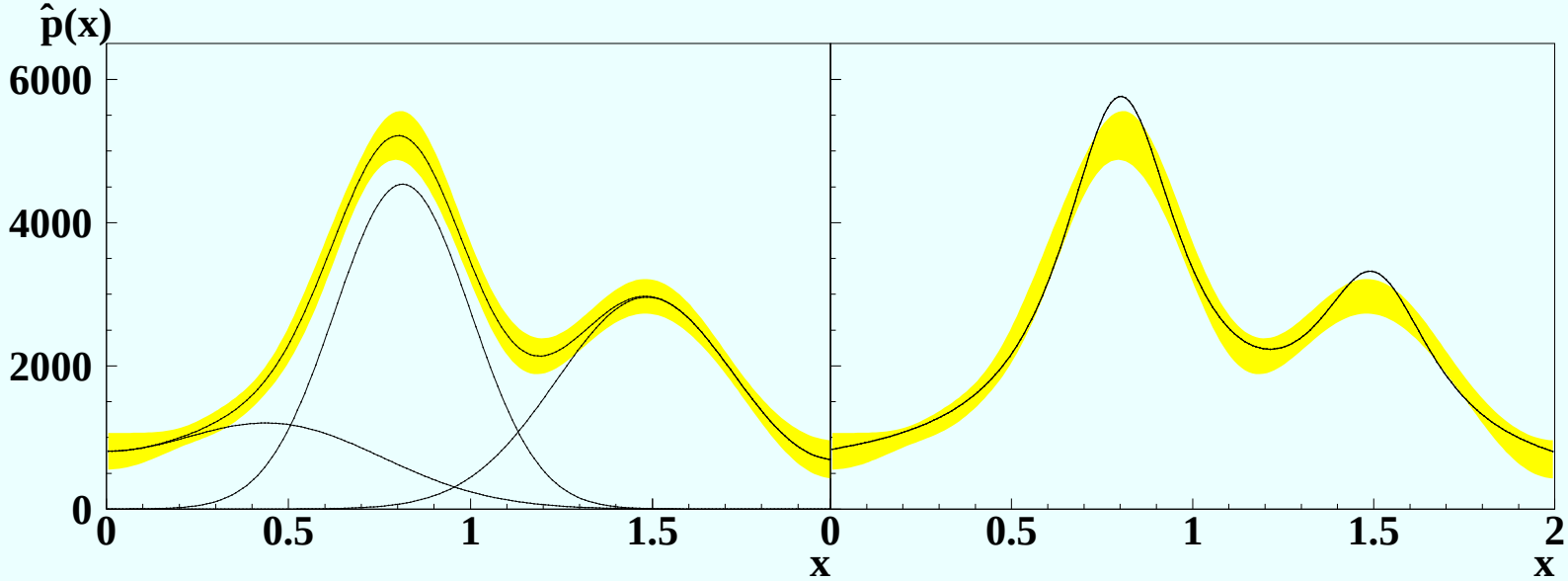


Figure: Components of the unfolded distribution and the unfolded distribution $\hat{p}(x)$ given by the sum of the components with $\pm 2\delta(x)$ interval (left) and the error band overlaid with the true distribution $p(x)$ (right).

vspace 0.5cm

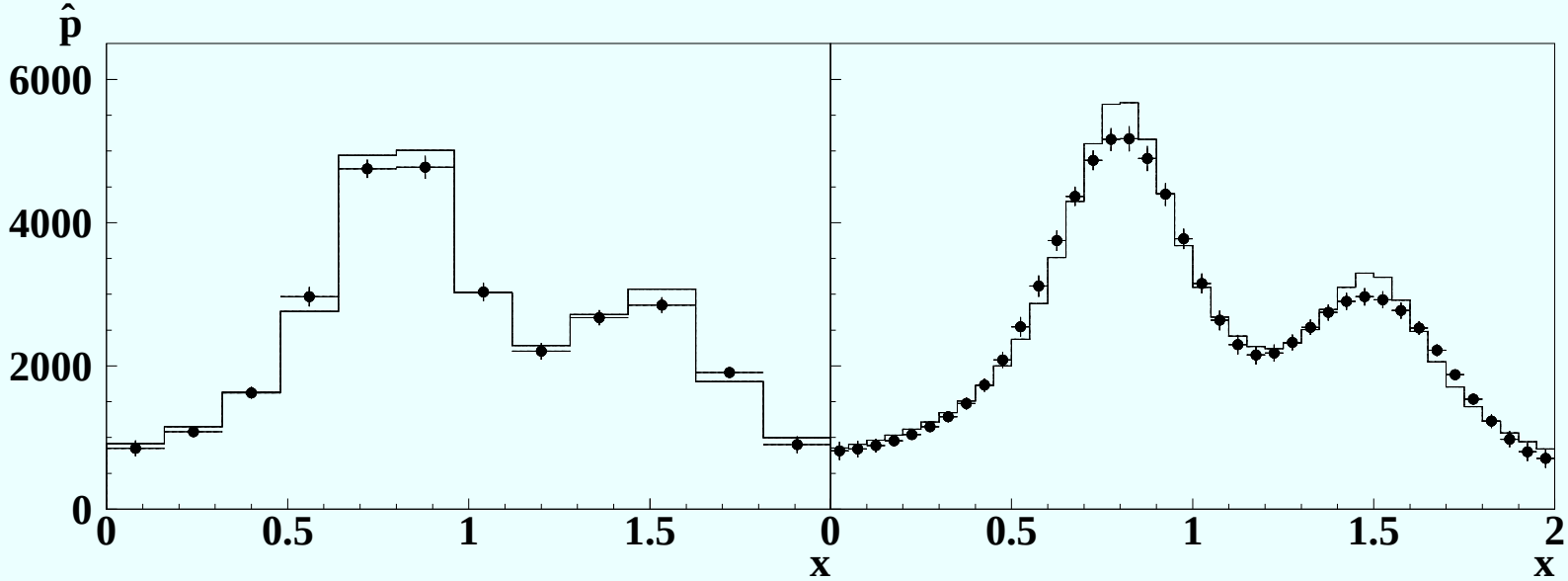


Figure: Unfolding results $\hat{p}_i$ for $m = 12$ (left) and $m = 40$ (right). The histogram show the true bin contents p_i.

vspace 0.5cm

Conclusions

- A new adaptive method for unfolding the true distribution from experimental data is presented.
- Proposed procedure uses Mixture Density Model for the representation of unknown true distribution and cross-validation approach to define optimal parameters of those model.
- Method does not need resolution function and acceptance in analytic form. To solve the problem of identification of the set-up, the Monte-Carlo sample is used, where each event represents the true and reconstructed parameters. Unfolding problem is solved simultaneously with identification problem.
- Proposed procedure can be used for solving a multidimensional problem.

vspace 0.5cm

Háskólinn á Akureyri
Max-Planck-Institut für Kernphysik

Notice:

A histogram representation for the unfolded distribution $\hat{p}(x)$ with m bins integrating over the x-intervals $[b_{i-1}, b_i]$, $i = 1, \ldots, m$ is obtained by

$$\hat{\boldsymbol{p}} = \mathbf{K}\,\hat{\boldsymbol{w}}, \tag{22}$$

where $\mathbf{K}$ is an $m \times k$ matrix with elements

$$\mathrm{K}_{ij} = \int_{b_{i-1}}^{b_i} K_j(x, x_j, \lambda_j)\, dx\ . \tag{23}$$

vspace 0.5cm

Example 2

The method is demonstrated for unfolding a steeply falling PDF. The true distribution, defined in the range $[0, +\infty)$, is

$$p(x) \propto xe^{-5x}. \tag{24}$$

Let us consider the variable $x = \sqrt{u^2 + v^2}$, where $u = x\cos(\phi)$ and $v = x\sin(\phi)$. The angle ϕ has uniformly distributed in $[0, 2\phi)$.
The reconstructed value $x' = \sqrt{u'^2 + v'^2}$, where u', v' are defined as independent random variables with normal distributions $\mathcal{N}(u, (0.5u)^2)$ and $\mathcal{N}(v, (0.5v)^2)$ respectively. The resolution function $R(x'|x)$, represent some generalization of the Rice distribution.

vspace 0.5cm

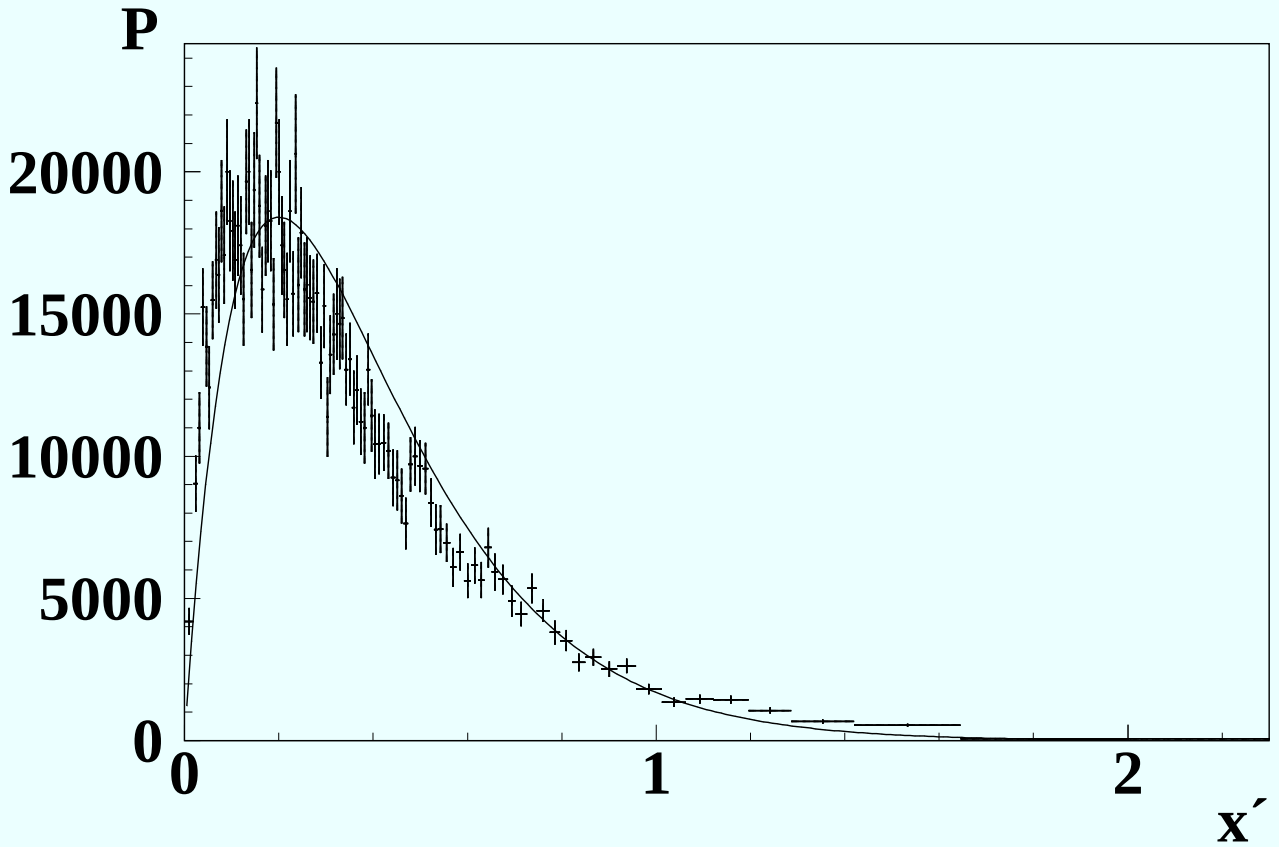


Figure: The measured distribution $P(x)$ (number of events divided on bin size). The true distribution $p(x)$ is shown as curve.

vspace 0.5cm

It is wise to transform true distribution to make it close to a gauss distribution and then use for the fit GMM model. The Box-Cox transformation can be applied in this case:

$$x^{\lceil\mu\rfloor} = \begin{cases} (x^{\mu} - 1)/\mu & \text{for } \mu \neq 0 \\ \ln x & \text{for } \mu = 0 \end{cases} \tag{25}$$

And then transform the result of fitting back, that is equivalent to represent the true distribution as mixture of distributions:

$$K(x, x_i, \lambda, \mu) = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{(x^{\lceil\mu\rfloor} - x_i^{\lceil\mu\rfloor})^2}{2\lambda^2}\right) x^{(\mu-1)} \tag{26}$$

vspace 0.5cm

The measured distribution obtained by simulating a sample of $N = 10000$ events.

A set of 400 PDFMs was used with positions x_i uniformly distributed over the interval $[0, 2.3]$ for the first step.

The cross-validation method with 5 folds was used.

We choose parameter $\mu = 0.25$, because transformed PDF has skewness close to 0 in this case.

vspace 0.5cm

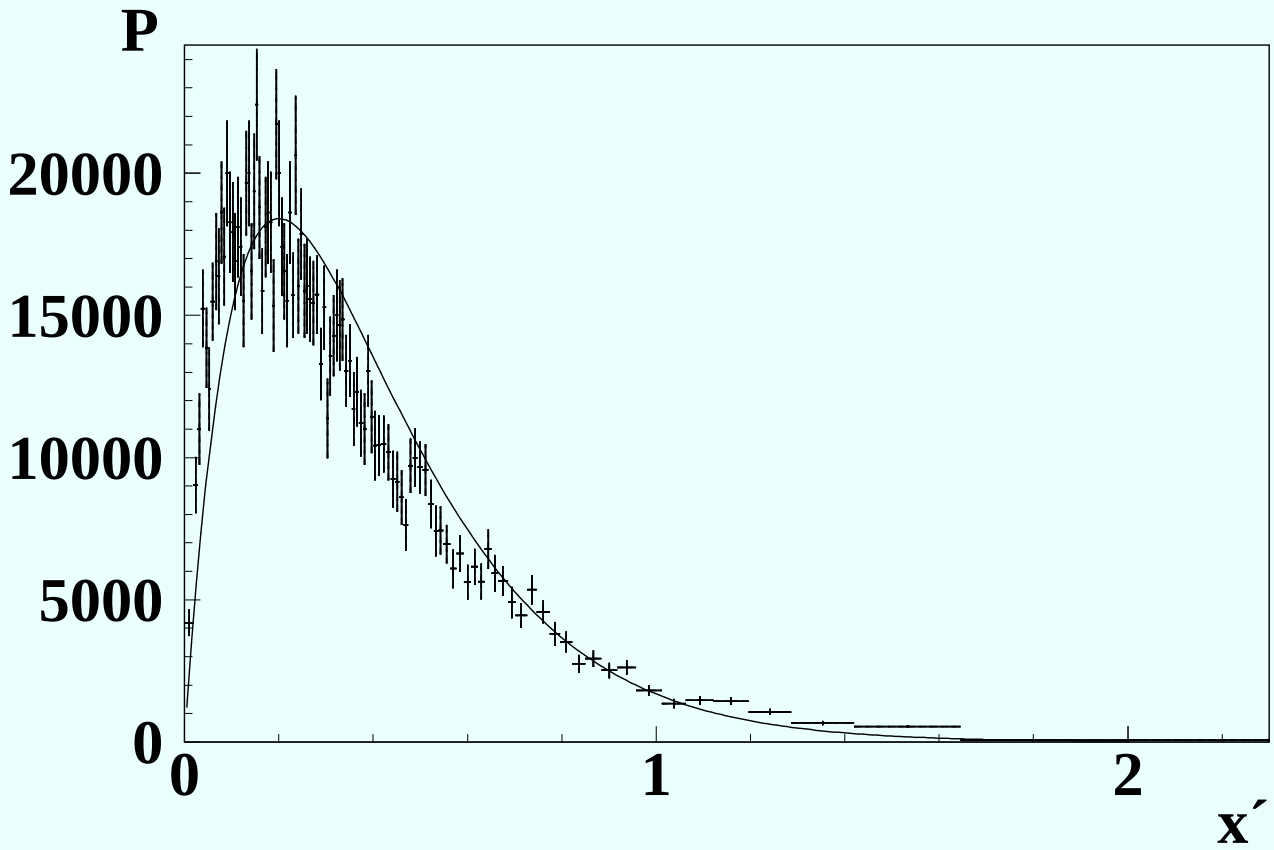


Figure: The measured distribution $P(x)$ (number of events divided on bin size). The true distribution $p(x)$ is shown as curve.

vspace 0.5cm

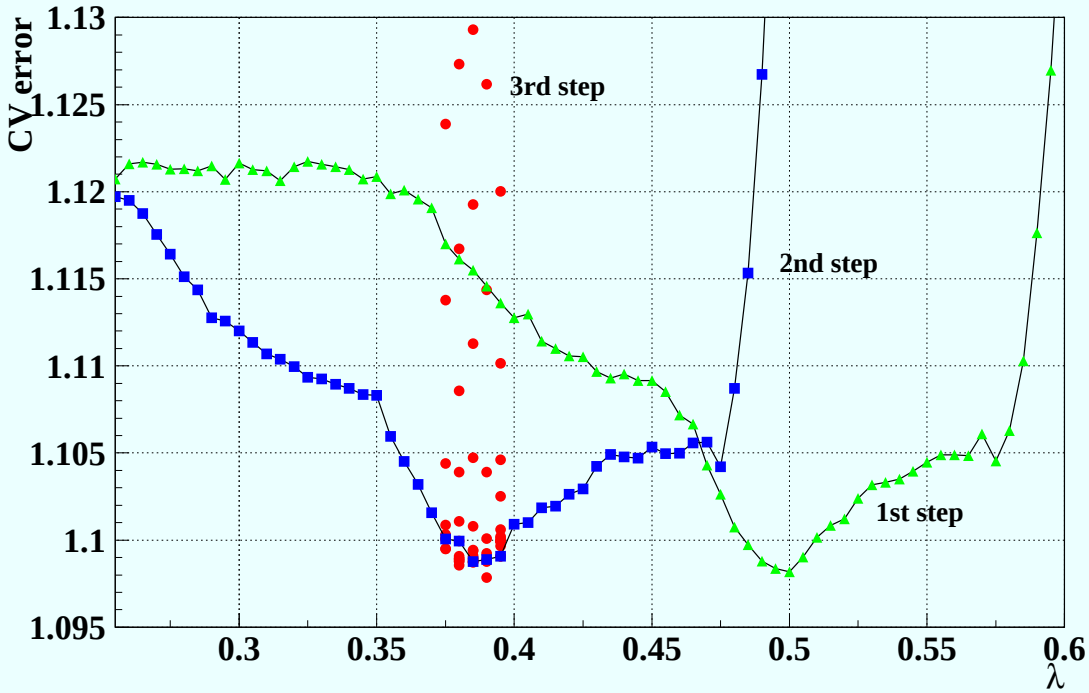


Figure: Cross Validation error for different values of λ

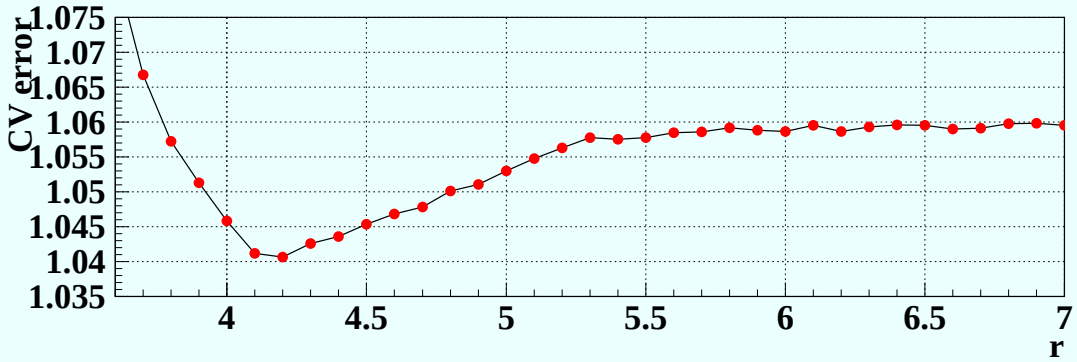


Figure: Cross Validation error for different values of r, $\lambda = 0.39$

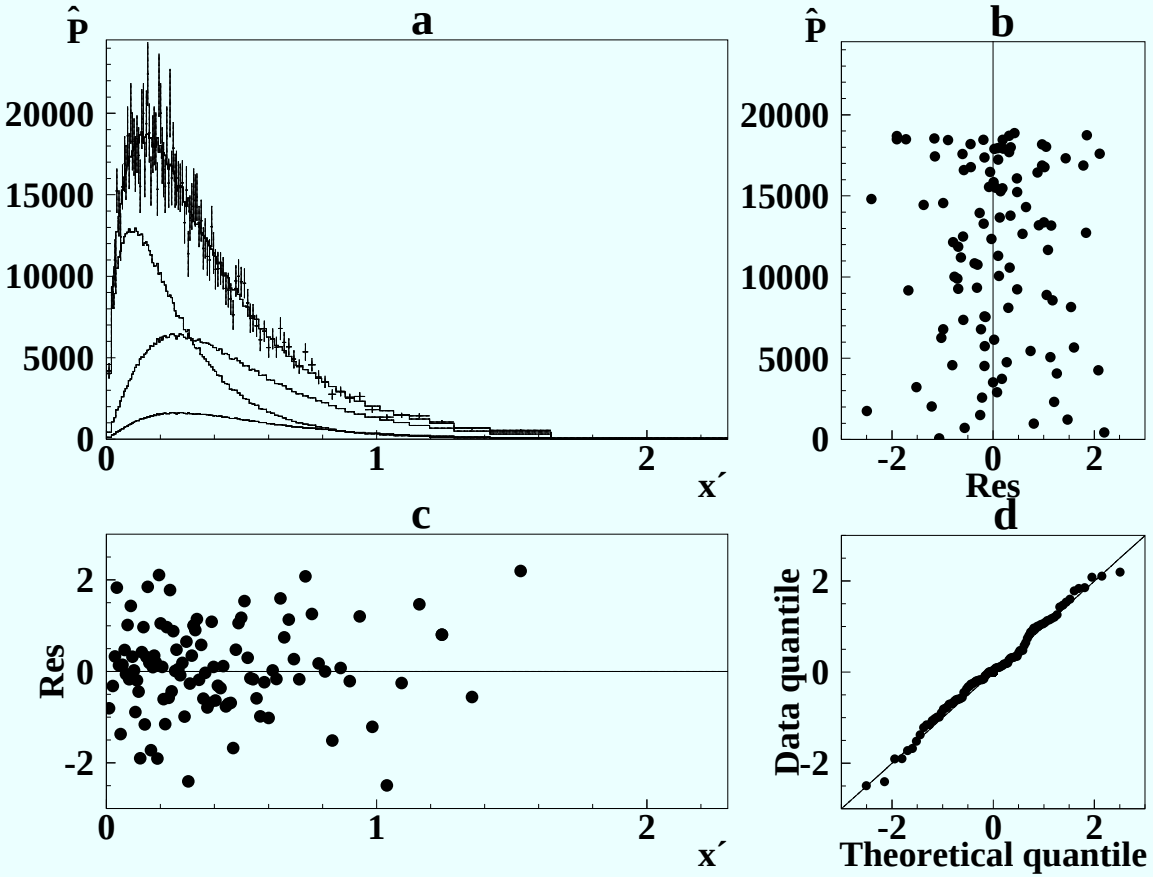


Figure: Illustration of the quality of the unfolding result. (a) folded PDFMs of the estimate of the true distribution compared to the measured distribution; (b) normalized residuals of the fit as a function of $\hat{\boldsymbol{P}}$; (c) normalized residuals as a function of x';
(d) quantile-quantile plot for the normalized residuals.

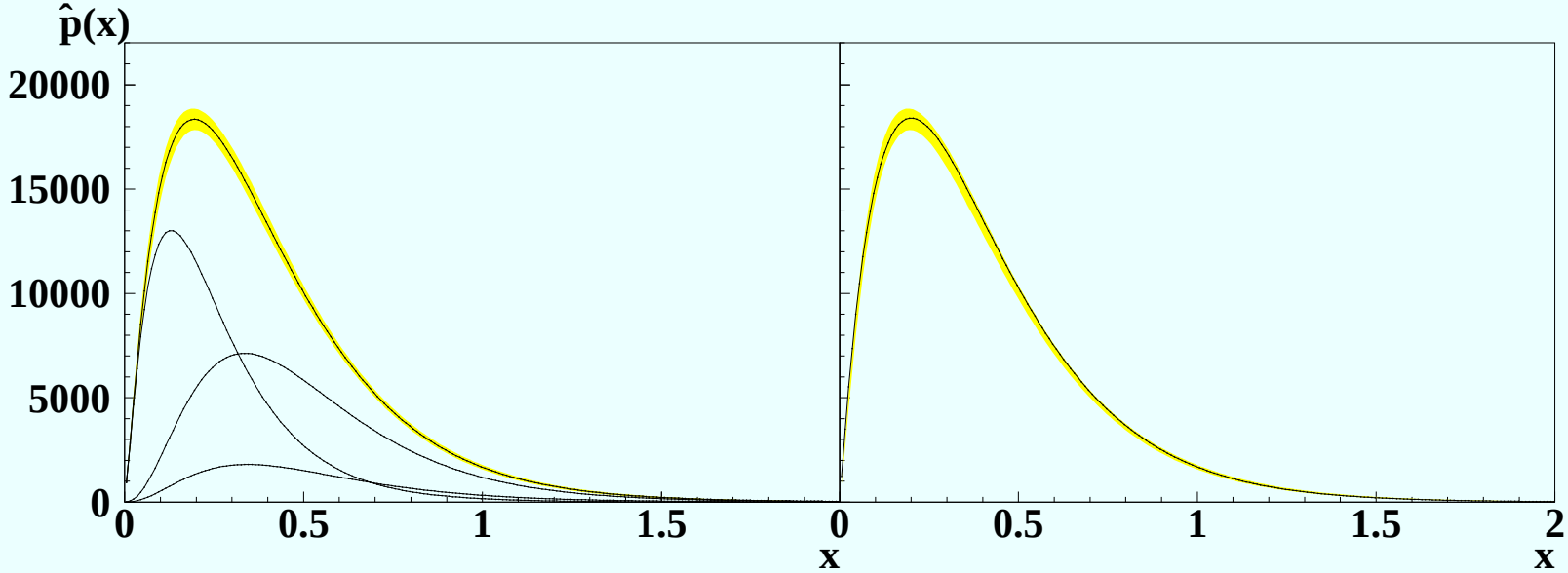


Figure: Components of the unfolded distribution and the unfolded distribution $\hat{p}(x)$ given by the sum of the components with $\pm 2\delta(x)$ interval (left) and the error band overlaid with the true distribution $p(x)$ (right).

vspace 0.5cm

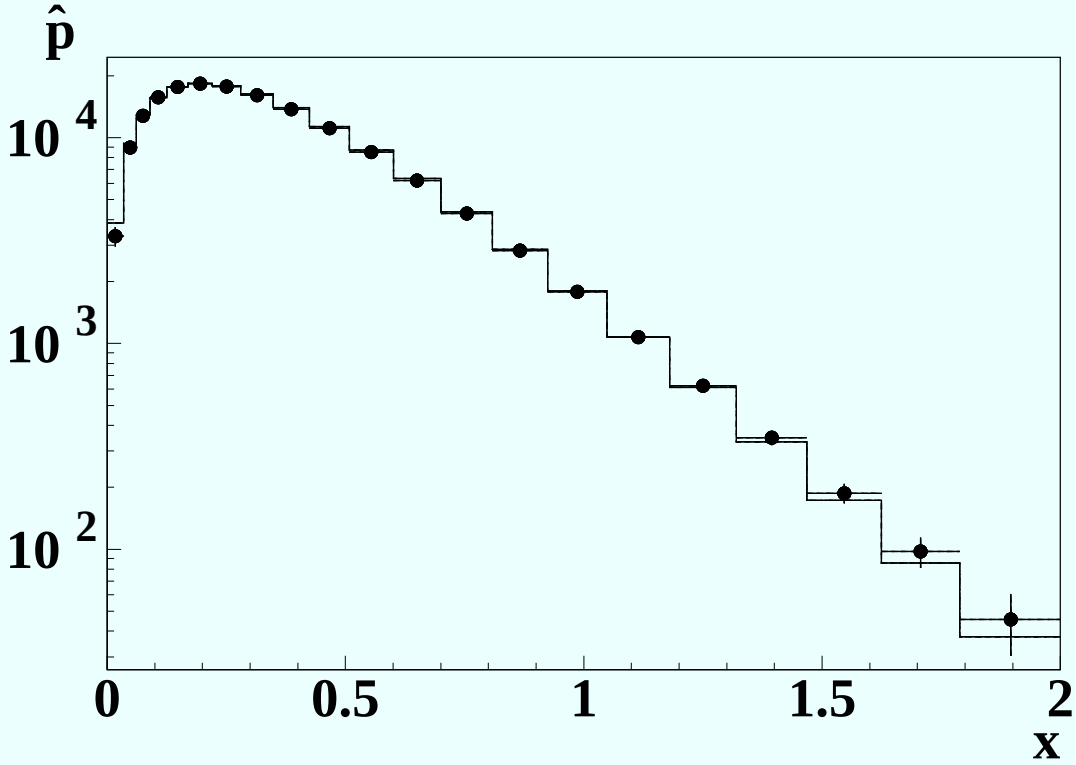


Figure: Unfolding results $\hat{p}_i$ for $m = 21$. The histogram show the true bin contents p_i.

vspace 0.5cm