Analysis of distorted measurements - parameter estimation and unfolding

Günter Zech
zech@physik.uni-siegen.de
Universität Siegen, Germany, July 26, 2016

Abstract:

Two different methods of parameter estimation are discussed to correct data for limited experimental resolution and acceptance losses: 1. The parameters are adjusted directly to the observed data and afterwards the introduced biases are corrected by a Monte Carlo simulation. 2. The experiment is simulated with a first guess $\theta_0$ of the parameter $\theta$. The simulated data are adjusted to variable parameter values $\theta$ by parameter depending weights. The parameters are inferred from a comparison of the weighted simulated data with the experimental data in form of histograms. If the statistical error of the simulation can be neglected, the parameters are determined in a Poisson likelihood fit. In the opposite case, a least square fit is performed where the sum of the weights in the histogram bins is approximated by a scaled Poisson distribution. This approximation is more precise than the usually used normal approximation.

In the non-parametric situation it is proposed to unfold the observed data in form of histograms with bins wide enough to suppress the excessive fluctuations caused by the noise. A likelihood fit is performed with the expectation maximization (EM) method. The resulting unfolded histogram and the error matrix can be used to combine data from different experiments and to fit parameters of distributions. The dependence of the results on the distribution used in the Monte Carlo simulation of the experiment is reduced to a negligible bias by a spline approximation of the distribution.

The common unfolding methods with explicit regularization are discussed and compared to each other. Approaches based on the eigenvector decomposition of the response matrix (equivalent to singular value decomposition), the iterative Richardson-Lucy (EM method) algorithm with early stopping, and three different penalty regularizations are investigated. To obtain an objective comparison of the methods, the smoothing parameter is optimized in each method such that the integrated square error (ISE) is minimal. Results depending on different distributions, experimental resolutions and event numbers are compared in a statistically significant number of toy experiments. The SVD method fails in some examples and the Tikhonov norm regularization performs less well than the remaining methods. Based on the ISE of the unfolded histogram, the best results are obtained with the iterative EM approach with early stopping. In addition to the histogram representation of the unfolded distribution, a superposition of basic spline functions is studied. Here, the curvature regularization and the EM method with early stopping produce promising results. The probability density estimates (PDEs) permit to visualize the true distribution and to discard predictions on a semi-quantitative level. Several methods to select the regularization parameter are discussed and a general method which is applicable to all unfolding schemes is proposed: The regularization parameter is adjusted iteratively in such a way that the mean integrated square error (MISE) is minimal. Because of the bias and the dependence of the error definition on the statistics, the applied regularization strength and the used smoothing method, the PDEs obtained by unfolding cannot be used for parameter inference and for quantitative goodness-of-fit tests. Therefore addition to the PDEs, unbiased unfolding results regularized by wide binning should be published. Parameter fits based on unfolded histograms with explicit regularization produce significantly biased results and larger uncertainties than direct fits of the parameters to the raw data.
## Contents

1 Introduction .................................................................................................................. 5

2 Parameter inference ....................................................................................................... 6
   2.1 Parameter correction method ..................................................................................... 8
   2.2 Weighting approach ................................................................................................. 10
      2.2.1 Negligible statistical error of the Monte Carlo simulation ................................. 10
      2.2.2 Data contaminated by background and correlated errors ................................. 18
      2.2.3 Including the statistical uncertainty of the simulation ..................................... 19
   2.3 Summary .................................................................................................................. 22

3 Discrete inverse problems and the response matrix ....................................................... 23
   3.1 Introduction and definition ....................................................................................... 23
      3.1.1 An inverse problem ......................................................................................... 23
      3.1.2 The histogram representation ........................................................................... 23
      3.1.3 Expansion of the true distribution .................................................................... 27
   3.2 The least square solution and the eigenvalue decomposition ................................... 27
      3.2.1 The least square solution .................................................................................. 27
      3.2.2 Eigenvector decomposition of the least square matrix ..................................... 28
      3.2.3 The effective number of parameters .................................................................. 33
   3.3 Summary .................................................................................................................. 35

4 Unfolding without explicit regularization ...................................................................... 35
   4.1 Introduction .............................................................................................................. 35
   4.2 The maximum likelihood approach ........................................................................ 36
   4.3 The Expectation-Maximization algorithm ............................................................... 36
   4.4 Diagonalizing the error matrix ................................................................................ 43
   4.5 Choice of the binning .............................................................................................. 43
   4.6 Dependence on the Monte Carlo input distribution .............................................. 46
      4.6.1 Uniform Monte Carlo distribution ................................................................. 46
      4.6.2 Spline approximation of the Monte Carlo input distribution ........................... 49
5 Unfolding with explicit regularization

5.1 General considerations

5.1.1 Regularization methods

5.1.2 Acceptance losses

5.1.3 Variable dependence and correlations

5.1.4 Measures of the unfolding quality in Monte Carlo experiments

5.1.5 Choice of the regularization strength

5.1.6 Error assignment to unfolded distributions

5.2 EM unfolding with early stopping

5.2.1 A few examples

5.2.2 Reduced iteration speed

5.2.3 Choice of the regularization strength

5.2.4 Introducing a final smoothing step

5.2.5 Dependence on the starting distribution

5.3 SVD based methods

5.3.1 Truncated SVD

5.3.2 Smooth truncation

5.3.3 Selective SVD

5.4 Penalty regularization

5.4.1 Curvature regularization

5.4.2 Entropy regularization

5.4.3 Tikhonov or norm regularization

5.5 Spline approximations

5.5.1 Curvature penalty

5.5.2 EM unfolding

5.6 Quality tests of the unfolding methods and comparison

5.6.1 The one-peak distribution

5.6.2 A two-peak distribution

5.6.3 A lifetime distribution

5.6.4 A pt distribution

5.7 Parameter estimation from unfolded histograms

5.8 Summary and recommendations

6 Summary of the summaries and a personal recommendation
6.0.1 Quadratic b-splines .......................................................... 108

References ................................................................................. 114
1 Introduction

The detectors of physics experiments are never perfect. They suffer from acceptance losses and from their finite resolution. As a consequence the distributions that we want to measure are distorted. A typical example is shown in Fig. 1.1 where an observed lifetime distribution is displayed.

The original distribution, we call it the *true distribution*, is smeared and events at the low end of the scale are lost. Our goal is to recover as much as possible the true distribution of the data and in this specific example, we want to extract the interesting parameter, the true lifetime.

In the past, until about 1990, in most experiments it was adequate to perform simple corrections to distorted distributions, but in the 90ties when structure functions were measured in electron proton collisions at the electron proton collider HERA, more sophisticated unfolding methods had to be applied. With new data from the Large Hadron Collider (LHC) at CERN the analysis of distorted data gained new interest. There exist by now a considerable number of publications on unfolding and parameter reconstruction. Some early publications [1, 2, 3, 4, 5, 6] after the pioneering work of Tikhonov [7] remained essentially unnoticed by particle physicists. Blobel was the first to promote professional unfolding in our field [8]. Since then, a large number of different approaches and studies were published. [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. A workshop was held at CERN [25] with interesting contributions but diverging proposals of the participants and an introduction by Lyons pointing to the essential problems of unfolding which we try to address in this report. An elementary introduction to unfolding is given in a note by Cowan [26] and a professional introduction to conventional unfolding with a the focus on singular value decomposition is presented in the book by Hansen [27].

![Fig. 1.1. Smeared lifetime distribution.](image-url)
Most authors from the particle physics community demonstrate the quality of their proposed methods with Monte Carlo studies, based on a single or very few data sets and selected distributions. It is not clear how well the obtained results and conclusions can be generalized and sometimes the choice of the free parameters of the models is somewhat arbitrary because they cannot be derived from first principles. The aim of this report is to summarize, test, extend and compare existing methods, but clearly not all aspects of unfolding can be covered. The author’s hope is that further systematic studies are stimulated by this report and that in this way a consensus among particle physicists can be reached on how distorted data should be analyzed.

The experimental situation is usually the following: We are given a sample of observations which we call events, drawn from a statistical distribution. Each event is characterized by a set of measured variables like energy, momentum, time etc. To simplify the following discussion, we limit ourselves to a single variable which we denote by $x$. The extension of the results to several variables is straightforward. To extract useful information from the sample, we have to understand the detector properties sufficiently well, such that we can simulate the expected distribution of observed events for a given undistorted true distribution.

Let us distinguish three different situations and issues:

1. We have of a parametric model. The true distribution from which the observed events are generated is given up to some unknown parameters. We want to infer the parameters.

2. We completely ignore the true distribution, but we want to prepare the data in such a way that they can be quantitatively compared to theoretical predictions and to results from other experiments. Furthermore, we require that results from different experiments can be combined. To this end, we parametrize the true distribution, usually in form of a histogram or by a spline approximation and we determine the corresponding parameters, i.e. the content of the histogram bins or the spline coefficients.

3. As in issue 2, we intend to estimate the completely unknown true distribution, but have the prior information that the distribution is smooth of which we want to take advantage of. The true distribution has to be parametrized and the parameters have to be estimated under constraints which correspond to our smoothness prejudices. Most unfolding procedures refer to this issue.

As we will see, issue 1 has a relative simple solution. If we have a parametric model where $f(x|\theta)$ is known up to unknown parameters $\theta$, we can estimate the parameters with least square (LS) or maximum likelihood (ML) methods. We compare the folded true distribution to the observed distribution and vary the parameter of the true distribution until the LS statistic is minimum or the likelihood is maximum. In some cases we can fit the parameters of $f(x|\theta)$ to the observed sample ignoring the experimental distortion and correct for the bias of the results by a Monte Carlo simulation.

The situation is similar for issue 2. We have to decide for the kind of parametrization and choose the number of parameters, i.e. the number of bins in case the result is presented in form of a histogram. This number has to be relatively small in order to avoid strong fluctuations, excessive correlations and huge diagonal errors. Another, but not very realistic possibility which permits a larger number of bins, would be to restrict the fluctuations by a smoothing algorithm which than has to be published together with the result. The same smoothing could then be applied to arbitrary predictions before they are compared to the experimental result. A quantitative comparison or a combination of data from different experiments would not be possible.

Issue 3 is what usually is meant with the notion of unfolding. It belongs to the field of parameter density estimation (PDE) [28, 29]. It is less well defined than the other two issues because there is no precise definition of smoothness. Smoothness is not invariant under transformations of the random variable and usually our prejudices of what is smooth depend on the problem that we have to solve. The result of unfolding a mass distribution will usually be incompatible with that obtained
from unfolding a mass squared distribution. Also the smoothness criteria may be different when we investigate a line spectrum superposed to a background or when a transverse momentum distribution in a particle experiment has to be unfolded. The smoothness constraints improve formally the precision of the results but introduce a bias. A compromise between precision and bias has to be found. Since the result and the size of the errors obtained from the adjustment depend crucially on the smoothness assumptions, the unfolding result is not suited for the inference of parameters of distributions. However it provides a semi-quantitative illustration of the true distribution and is normally closer to the true distribution than a simple parametrization. It can be used to discard theoretical concepts if the discrepancy between the unfolding solution and the prediction is large and occasionally the results may be useful as input for simulation, for instance of structure functions. Due to the complexity of the unfolding problematic, the corresponding chapter in this report is much longer than the others.

Some colleagues believe that the issues 2 and 3 can be solved with a single method. However, it is not clear how the smoothing parameters can be fixed and how the systematic uncertainties in the parameter estimation can be handled.

Unfolding in higher dimensions suffers from the *curse of dimensionality*: The bins of multi-dimensional histograms are often sparsely populated. This problem can partially be solved with binning-free unfolding methods. A modest, explorative study is presented in [30].

A technical remark: In many of the figures of this report the titles of the axes and even scales have been omitted to save space where they are obvious or not necessary for the understanding of the intended message.

Some parts of this report have been copied with minor modifications from the book by Bohm and myself [30].

## 2 Parameter inference

In particle physics we are in the lucky situation that in most cases we have a theoretical description of the data that we collect. The reason is that the experiments are usually designed with the goal to test a prediction or to measure parameters of it. We may want, for instance, to determine the lifetime of a particle from an exponential distribution or the mass and width of a particle from a Breit-Wigner distribution.

One might think, that first of all, one should unfold the observed distribution to get rid of the experimental defects and then pursue with the analysis. However, this is not a good idea, because unfolding is not straightforward and it is accompanied by a loss of information. It is much better to fold the prediction and to compare the folded prediction to the observed data. In this way we can avoid some of the approximations that are necessary in unfolding methods and we do not have to care about the oscillations that occur in unfolding procedures due to the statistical fluctuations of the data.

In [21] a so-called parametric unfolding method is proposed. This approach has the disadvantage that a response matrix has to be constructed which depends on the distribution chosen to simulate the smearing. This dependence can be avoided by iteration but this is not necessary in the much simpler weighting procedure described in Sect. 2.2.
Fig. 2.1. Fit of a linear distribution. Left hand: Histogram of the data and shape of the true and the

distorted distributions. Right hand: Transition from the observed parameter to its true value.

2.1 Parameter correction method

Often the true distribution is only slightly distorted by the measurement. Then we can initially

neglect the experimental effects and fit the parameter we are interested in. We obtain a biased

maximum likelihood or least square estimate $\hat{\theta}'$ and an uncertainty $\delta \theta'$. The bias is then estimated

by a Monte Carlo simulation based on a value $\theta$ close to the true value or on $\hat{\theta}'$. The bias is usually

within the uncertainties independent of the value chosen in the simulation. In more than 99% of

all measurements in particle physics this simple method is applied or unnecessary if the bias is

negligible.

If the distortions are large, we have to generate the true distribution for a few values of $\theta$

and simulate the estimation of $\theta'$ to obtain the relation $\theta(\theta')$ between the parameter of interest $\theta$

and the observed quantity $\theta'$. The relation $\theta(\theta')$ can usually be taken to be linear in the vicinity of $\hat{\theta}'$

and then the choice of two values of $\theta$ is enough. The method should become clear in the following

examples.

Example 1. Polarization measurements of hyperons often lead to fits of linear cosine distributions. In

Fig. 2.1 such a distribution is displayed. The dashed line shows a p.d.f. $f(x) = 0.5 + 1.1x$, where $x$ is

confined to the interval $-1 \leq x \leq 1$. The folded version of this distribution, with a kernel following

a normal distribution with standard deviation $\sigma_s = 0.2$,

$$g(x) = \int_{-\infty}^{\infty} \mathcal{N}(x'|x, \sigma_s)f(x')dx',$$

$$\mathcal{N}(x'|x, \sigma_s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left[-\frac{(x' - x)^2}{2\sigma_s^2}\right],$$

extends to regions outside the interval borders of the true distribution. A histogram from a sample

of 100 events, $x_1, x_2, ..., x_{100}$ generated according to the folded distribution is displayed in the same

figure. A maximum likelihood fit of the parameter $\alpha$ to the observed data with the log-likelihood function

$$\ln L(\alpha) = \sum_{i=1}^{100} (0.5 + \alpha x_i)$$
Fig. 2.2. Experimental lifetime distribution. The insert indicates the transition from the observed to the corrected lifetime.

yields the value $\hat{\alpha}' = 0.065^{+0.080}_{-0.082}$. To correct this number for the smearing effect, several samples of $10^5$ events each, are generated with different values of $\alpha$. From the corresponding fitted values a within statistics linear relation $\alpha' = 0.90\alpha$ is derived as shown at the right-hand side of Fig. 2.1. This relation is then used to correct for the bias of $\alpha'$. The result is $\hat{\alpha} = \hat{\alpha}'/0.90 = 0.072^{+0.089}_{-0.091}$. The factor 0.9 which relates $\alpha$ and $\alpha'$ indicates that the distortion of the measurement entails a loss in precision of 10%.

**Example 2.** In Fig. 2.2 an observed distorted lifetime distribution is depicted. The sample mean $\bar{t}$ of a sample of $N$ undistorted exponentially distributed lifetimes $t_i$, is a sufficient estimator of the mean lifetime $\tau$. It contains the full information related to the parameter $\tau$, the mean lifetime. In case the distribution is distorted by resolution and acceptance effects, the mean value

$$\bar{t}' = \frac{\sum t_i'}{N}$$

of the distorted sample $t_i'$ will usually still contain almost the full information relative to the mean life $\tau$. The relation $\tau(\bar{t})$ between $\tau$ and its approximation $\bar{t}$ (see insert of Fig. 2.2) is generated by a Monte Carlo simulation. The uncertainty $\delta\tau$ is obtained by error propagation from the uncertainty $\delta\bar{t}$ of $\bar{t}$,

$$\delta\bar{t}^2 = \frac{(\bar{t}'^2 - \bar{t}^2)}{N-1},$$

$$\bar{t}'^2 = \frac{1}{N} \sum t_i'^2$$

using the Monte Carlo relation $\tau(\bar{t})$. 
The correction approach has several advantages:

- It is not necessary to histogram the observations. A likelihood fit with individual observations can be performed.
- Problems due to small event numbers for bins in a multivariate space are avoided.
- It is robust, simple and requires little computing time if a sufficient statistic exists. It is ideal for online applications.
- All approximations are automatically corrected for by the simulation.

As we have to perform a likelihood or a $\chi^2$ fit, in the cases where we do not have a sufficient statistic, we occasionally may run into the following problem: The data can lay outside the range covered by the undistorted p.d.f. In fact this also happens in the first example of this section. It did not cause any trouble, we just had to extrapolate the linear p.d.f. $f(x)$ to values below $-1$ and above $+1$. However, if the distribution is steep, some observed data values could correspond to negative function values where the log-likelihood is not defined. The problem can usually be solved by a linear transformation of the observed variable in such a way that it is covered by the range of the variable in which the p.d.f. is defined. In the example depicted in Fig. 2.1 this is $[-1, 1]$. Here dividing $x$ by 2, values outside the interval $[-1, 1]$ are excluded. The result obtained above and its error remain unchanged. The Monte Carlo simulation of the analysis procedure automatically corrects for the scaling.

Our examples concerned simple, smooth distributions. Complex, multi-modal distributions with large distortions cannot always be handled with the correction method without a sizable loss in precision.

### 2.2 Weighting approach

Now we turn to the standard method [31] that should be used in all cases where the resolution effects are so large that the simple correction method does not produce precise results. Now, the experimental and the simulated data are compared in form of histograms.

#### 2.2.1 Negligible statistical error of the Monte Carlo simulation

If the data sample is not extremely large, we can generate enough Monte Carlo events such that their statistical error can be neglected.

We compare a data histogram with $B$ bins and bin contents $d_i$ to a Monte Carlo generated histogram with bin contents $t_i$. (We use the letters $d$ and $t$ to denote data and theory.) To produce the Monte Carlo histogram, events are generated according to the p.d.f. $f(x|\theta)$ and the detector response is simulated. The corresponding observed variables $x'$ are then histogrammed. We get $d_i$ and a prediction $c t_i(\theta)$ where $c$ is the normalization constant. Usually the parametric model does not predict the number of observed events but only the shape of the distribution. Then the normalization $c$ is a free parameter. We assume that the number of events $d_i$ in a bin $i$ is Poisson distributed with the expected value equal to $c t_i$, $d_i \sim P(d_i|c t_i)$ with the abbreviation $P(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$.

#### The likelihood function

From the Poisson distribution for bin $i$
2.2 Weighting approach

\[ \mathcal{P}(d_i|ct_i) = \frac{e^{-ct_i}(ct_i)^{d_i}}{d_i!} \]

we derive the likelihood and its logarithm. The probabilities for the different bins are independent. We get

\[ L(c, \theta) = \prod_{i=1}^{B} \frac{e^{-ct_i}(ct_i)^{d_i}}{d_i!}, \quad (2.1) \]

\[ \ln L(c, \theta) = \sum_{i=1}^{B} \left[ -ct_i + d_i \ln(ct_i) \right] + \text{const}. \quad (2.2) \]

The parameter dependence is hidden in the numbers \( t_i(\theta) \) and the estimate \( \hat{\theta} \) is obtained by maximizing the log-likelihood with respect to \( \theta \). The errors \( \delta_- \), \( \delta_+ \) are derived in the usual way from the change of the log-likelihood by half a unit: \( \ln L(\hat{\theta}) - 1/2 = \ln L(\hat{\theta} - \delta_-) = \ln L(\hat{\theta} + \delta_+) \), or, if the statistics is high enough from the second derivative of the log-likelihood function at its maximum: \( \delta^2 = \left[ \frac{d^2 \ln L}{d \theta^2} \right]^{-1}_{\hat{\theta}} \).

An obvious estimate of the parameter \( c \) is the ratio of the total number \( N = \sum d_i \) of observed events and the total number \( M = \sum t_i \) of simulated events, \( \hat{c} = N/M \). This is also the maximum likelihood estimate: Deriving \( \ln L \) with respect to \( c \) and setting the derivative equal to zero,

\[ \frac{d \ln L}{dc} = \sum_{i=1}^{B} (-t_i + d_i/c) = 0, \]

we reproduce the consistent result \( \hat{c} = \sum d_i/\sum t_i = N/M \).

There is an alternative formulation of the problem: We can calculate the probabilities \( \varepsilon_i = t_i/\sum t_i \) for an event to fall into bin \( i \) and describe the data histogram by a multinomial distribution

\[ M_{\varepsilon_1, \ldots, \varepsilon_B}(d_1, \ldots, d_B) = \frac{N!}{\prod_{i=1}^{B} d_i!} \]

with the constraint \( \sum d_i = N \). Here a normalization is obsolete. The multinomial formulation is equivalent to the multi-Poisson formulation with normalization fixed to \( c = N/M \). It is not recommended to follow the multinomial way because the errors of the different numbers \( d_i \) become correlated and then the calculations are quite clumsy or unnecessary approximations have to be made.

**Variation of the parameter by weighting the events**

To estimate the parameter of interest \( \theta \) we have to maximize the log-likelihood (2.2), varying \( \theta \). At first sight one might think that for each new choice of \( \theta \), the complete Monte Carlo simulation has to be repeated. To proceed in this way does not work because it would require a huge amount of computer time\(^1\). The parameter change is implemented by re-weighting the Monte Carlo events in the observed histogram. To each event \( j \) in bin \( i \) with the observed variable \( x'_{ij} \), generated with the p.d.f. \( f(x_{ij}|\theta_0) \) we associate the weight

\[^1\text{The reason is related to the fact that a repetition of the simulation induces a large statistical modification of the likelihood or of } \chi^2 \text{ independent of a parameter change.}\]
Parameter inference

The weighted event variables \( x \) follow the p.d.f. \( f(x) \) and the weighted observed variables follow the smeared distribution which is compared to the observed data distribution. To compute the weight, we have to remember for each Monte Carlo event the true variable value \( x_{ij} \). The sum \( \sum_{j=1}^{m_i} w_{ij} \) of the weights of the \( m_i \) entries in bin \( i \) corresponds to the prediction \( t_i(\theta) \):

\[
t_i(\theta) = \sum_{j=1}^{m_i} w_{ij} = \sum_{j=1}^{m_i} f(x_{ij}|\theta)/f(x_{ij}|\theta_0) .
\]

The values \( t_i(\theta) \) have to be inserted into (2.2). The normalization \( c \) can be set, \( c = N/\sum_{ij} w_{ij} \) or left free in the fit. The second possibility is simpler. To simplify the formulas we introduce the mean value \( \bar{w}_i \) of the \( m_i \) weights of the bin \( i \):

\[
\bar{w}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} w_{ij} .
\]

The corresponding log-likelihood is

\[
\ln L(\theta) = \sum_{i=1}^{B} \left[ -cm_i \bar{w}_i(\theta) + d_i \ln(cm_i \bar{w}_i(\theta)) \right] + \text{const.} .
\]

Example 3. We consider a superposition of a narrow Gaussian with a uniform background distribution in the interval \([0, 1]\). The free parameters are the mean \( \mu \), the standard deviation \( \sigma \) of the normal distribution and the background fraction \( \rho \).

\[
f(x) = \rho + (1 - \rho) \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]
\]

A “data” sample is generated with the parameters \( \mu_d = 0.5, \sigma_d = 0.05 \) and \( \rho_d = 0.5 \), smeared according to a normal distribution with standard deviation \( \sigma_s \) and histogrammed into 20 bins. A
Fig. 2.4. Observed events (histogram), fitted distribution (curve) and corresponding smeared and normalized histogram (dots). The two plots correspond to response functions with standard deviations of 0.04 and 0.06, respectively.

The following example is more involved.

Example 4. The function to be fitted is a superposition of an exponential and two normal distributions,
\[
f(x|\nu_1, \gamma, \mu_1, \sigma_1, \mu_2, \sigma_2) = \nu_1 \gamma \exp(-\gamma x) + \nu_1 \frac{1}{\sqrt{2\pi} \sigma_1} \exp \left[ -\frac{(x - \mu_1)^2}{2\sigma_1^2} \right] \\
+ (1 - \nu_1) \frac{1}{\sqrt{2\pi} \sigma_2} \exp \left[ -\frac{(x - \mu_2)^2}{2\sigma_2^2} \right],
\]

with seven parameters, \( \gamma, \mu_1, \sigma_1, \mu_2, \sigma_2, \nu_1, \nu_2 \). The data sample consists of 5000 events that are generated with the parameters quoted in the second column of the following table. The Monte Carlo sample contains 50 times more events. Events are generated in the interval \([-0.1, 1.1]\) but in the fit only events observed in the range \([0, 1]\) are considered. Two different normally distributed smearing functions with standard deviations \( \sigma_s = 0.04 \) and \( \sigma_s = 0.06 \) were applied. The results of the fits are summarized in the following table and shown in Fig. 2.4. While the uncertainty of the slope parameter of the exponential contribution changes by a negligible amount with the increased smearing, the resolution of the parameters of the two normal contributions change by about a factor two. Despite the large distortion of the original distribution in the case \( \sigma_s = 0.06 \), the fit is able to reproduce the true shape quite well.

| parameter | nominal | \( \sigma_s = 0.04 \) | \( \sigma_s = 0.06 \) |
|-----------|---------|----------------|----------------|
| \( \gamma \) | 1.00    | 1.072 ± 0.093 | 1.041 ± 0.096 |
| \( \mu_1 \)  | 0.40    | 0.392 ± 0.007 | 0.390 ± 0.015 |
| \( \sigma_1 \) | 0.05    | 0.045 ± 0.008 | 0.040 ± 0.019 |
| \( \mu_2 \)  | 0.60    | 0.597 ± 0.007 | 0.596 ± 0.014 |
| \( \sigma_2 \) | 0.05    | 0.058 ± 0.009 | 0.055 ± 0.015 |
| \( \nu_1 \)  | 0.70    | 0.677 ± 0.020 | 0.691 ± 0.022 |
| \( \nu_2 \)  | 0.15    | 0.143 ± 0.015 | 0.134 ± 0.025 |

### The least square formulation

If the observed distribution is not described by Poisson statistics, we cannot apply the maximum likelihood method and have to turn to the least square formalism which is the second best choice.

We assume again that the number of simulated events \( M \) is much larger than the number of observed events \( N \). The number of events in the histogram bins are assumed to be Poisson distributed, with the expected value \( c \mu_i \). We form a test quantity \( \chi^2 \),

\[
\chi^2 = \sum_{i=1}^{B} \frac{(d_i - c \mu_i)^2}{\delta_i^2},
\]

where \( \delta_i \) is the expected uncertainty of \( d_i - c \mu_i \) under the null hypothesis that the data are described by the prediction, i.e., that the expected value of the bracket equals zero, \( E(d_i - c \mu_i) = 0 \). With \( t_i \gg d_i \) we can neglect the statistical error of the simulation and in the denominator remains the expected uncertainty squared of \( d_i \) which for the Poisson distribution is the square root of the expected value \( c \mu_i \) of \( d_i \). We obtain:

\[
\chi^2 = \sum_{i=1}^{B} \frac{(d_i - c \mu_i \bar{w}_i)^2}{c \mu_i \bar{w}_i}.
\]

Remark that approximating \( \delta_i^2 \) simply by \( d_i \) is inconsistent and biases the results.

If the measurements are not Poisson distributed, the corresponding estimates of the errors \( \delta_i \) have to be inserted in (2.7).

Setting \( d \chi^2/dc = 0 \), we get the estimate \( \hat{c} \):
2.2 Weighting approach

\[
\hat{c} = \left[ \frac{1}{B} \sum_{i=1}^{B} d_i^2 \right]^{1/2}.
\]

Remark that this estimate differs from the MLE and is in disagreement with the corresponding estimate from multinomial formulation of the problem, however the difference is negligible in most cases. It does not really matter whether we fix the normalization to \( c = N/\Sigma_i m_i \bar{w}_i \) or leave it as a free parameter in the fit.

For large event numbers \( d_i \) the Poisson distribution can be approximated by a normal distribution and consequently \((d_i - ct_i)/\sqrt{ct_i}\) is normally distributed with variance equal one. Then our test quantity \( \chi^2 \) follows a \( \chi^2 \) distribution with \( B - P \) degrees of freedom, where \( P \) is the number of free parameters in the fit.

Bins with small event numbers are problematic, because the Poisson errors are strongly asymmetric and LSFs are optimal if the distributions can be approximated by normal distributions. Therefore for low statistics experiments one should select the bin width as wide as allowed by the band width of the distribution. The variation of the resolution with the bin width is estimated in Appendix 2 for a Gaussian peak.

**Parameter estimation in experiments with a large number of events**

Statistical problems decrease with increasing event numbers, but computational requirements increase. The numerical minimum search that is required to estimate the wanted parameters can become quite slow. It may happen that we need of the order of \( 10^6 \) or more simulated events. This means that, for say \( 10^4 \) changes of a parameter value during the extremum search that \( 10^{10} \) weights have to be computed. This is feasible, but we may want to speed up the fitting procedure.
The individual weighting of events can be avoided if the parameters appear in factors depending on the parameters only:

\[ f(x|\theta) = h_1(\theta)f_1(x) + h_2(\theta)f_2(x) + \ldots + h_n(\theta)f_n(x) \]  

(2.9)

Then we can simulate the smeared versions \( f'_i(x') \) to \( f'_n(x') \) of \( f_1(x) \) to \( f_n(x) \) and compare the observed histogram to the superposition of the histograms \( t(1) \) to \( t(n) \) of the smeared functions. We replace in (2.7) \( t_i \) by the sum \( t_i = h_1(\theta)t_i^{(1)} + h_2(\theta)t_i^{(2)} + \ldots + h_n(\theta)t_i^{(n)} \). Since we do not need to compute weights for individual events, the minimum search is accelerated drastically.

To illustrate the method we choose a simple example.

**Example 5.** We fit a straight line. The p.d.f. be \( f(x) = 1 - \theta/2 + \theta x \), \( 0 \leq x \leq 1 \). With \( f_1 = 1 \) and \( f_2 = x \) we have \( f(x) = (1 - \theta/2)f_1(x) + \theta f_2(x) \). We generate Monte Carlo events uniformly distributed in \( x \) and collect the smeared observed values \( x' \) in a histogram \( t(1) \). There may be acceptance losses. We proceed in the same way with \( f_2(x) \) and create a histogram \( t(2) \). Our prediction for \( d \) is \( c[(1 - \theta/2)t^{(1)}_i + \theta t^{(2)}_i] \) with the normalization factor \( c \). The \( \chi^2 \) expression becomes:

\[ \chi^2 = \sum_{i=1}^B \left[ d_i - c \left( (1 - \theta/2)t^{(1)}_i + \theta t^{(2)}_i \right) \right]^2 \left[ (1 - \theta/2)t^{(1)}_i + \theta t^{(2)}_i \right]. \]

Fig. 2.5 shows the results from a numerical example. The experimental data are simulated with 5000 events and a slope of \( \theta = 1 \). The histograms \( t^{(1)} \) and \( t^{(2)} \) are derived from \( 10^6 \) Monte Carlo events. The Gaussian resolution is 0.1. Fig. 2.5a) and b) are simulations of the observed uniform and the observed linear Monte Carlo histograms. A superposition of the two histograms is fitted to the observed distribution in Fig. 2.5. The fit result is \( \theta = 0.955 \pm 0.033 \).

In the superposition (2.9) the functions \( f_i(x) \) have to be positive integrable. For example, a linear distribution of the cosine \( z \) of the polar angle is described by \( f(z|\theta) = (1 + \theta z)/2 \), \(-1 \leq z \leq 1\). To avoid negative probabilities, we express the p.d.f. by \( f(x|\theta) = (1 + \theta)f_1 + (1 - \theta)f_2 \) with \( f_1 = (1 + z)/4 \), \( f_2 = (1 - z)/4 \).

Usually the p.d.f.s are not of the simple form (2.9). Then, if the parameter is known to be close to \( \theta_0 \), we can use a Taylor expansion of \( f(x|\theta) \) at \( \theta_0 \) in powers of the difference \( \Delta = \theta - \theta_0 \) with respect to the parameter at some preliminary estimate \( \theta_0 \):

\[ f(x|\theta) = f(x|\theta_0) + \frac{\Delta f(x|\theta)}{\Delta \theta} |_{\theta_0} + \frac{\Delta^2 f(x|\theta)}{2!} |_{\theta_0} + \ldots \]  

(2.10)

\[ = f(x|\theta_0) \left\{ 1 + \frac{\Delta}{f(x|\theta_0)} \frac{df(x|\theta)}{d\theta} |_{\theta_0} + \frac{\Delta^2}{2!} \frac{df^2(x|\theta)}{d\theta^2} |_{\theta_0} + \ldots \right\}. \]  

(2.11)

The powers of the new parameter \( \Delta \) factorize. We could now proceed as above and generate separately Monte Carlo events for each summand, but it is more economic to generate a single sample following \( f(x|\theta_0) \) and to apply weights. Let us assume that \( \Delta \) is small and that the expansion can be cut after the third term. The Monte Carlo events are collected into histograms of the observed values of the variable \( x' \) filled into the histogram \( t^{(0)} \). Weighting each entry by \( w_1(x) \) and \( w_2(x) \),

\[ w_1(x) = \frac{1}{f(x|\theta_0)} \frac{df(x|\theta)}{d\theta} |_{\theta_0} \]  

(2.12)

\[ w_2(x) = \frac{1}{2f(x|\theta_0)} \frac{df^2(x|\theta)}{d\theta^2} |_{\theta_0}. \]  

(2.13)
we generate the histograms \( t^{(1)} \) and \( t^{(2)} \).

The parameter inference of \( \Delta \) is performed by comparing the experimental histogram bins contents \( d_i \) to \( t_i = c(t_i^{(0)} + \Delta t_i^{(1)} + \Delta^2 t_i^{(2)}) \):

\[
\chi^2 = \sum_{i=1}^{B} \frac{(d_i - c t_i)^2}{c t_i} .
\]

(2.14)

In many cases the quadratic term can be omitted. In other situations it might be necessary to iterate the procedure.

This method works only if the number of Monte Carlo events is high enough to neglect its uncertainty with respect to that of the experimental data.

To illustrate the method, we consider a lifetime measurement:

**Example 6.** We expand the p.d.f.

\[
f(x|\gamma) = \gamma e^{-\gamma x}
\]

(2.15)

into a Taylor expansion at \( \gamma_0 \) which is a first guess of the decay rate \( \gamma \):

\[
f(t|\gamma) = \gamma_0 e^{-\gamma_0 x} \left\{ 1 + \frac{\Delta \gamma}{\gamma_0} (1 - \gamma_0 x) + \frac{\Delta^2 \gamma}{\gamma_0^2} (-\gamma_0 x + \frac{\gamma_0^2 x^2}{2} ) + \cdots \right\} .
\]

(2.16)
The Monte Carlo simulation follows the distribution \( f_0 = \gamma_0 e^{-\gamma_0 x} \). Weighting the events by \((1/\gamma_0 - x)\) and \((-x/\gamma_0 + x^2/2)\), we obtain the distributions \( f_1 = (1 - \gamma_0 x)e^{-\gamma_0 x} \), \( f_2 = (-x + \gamma_0 x^2/2)e^{-\gamma_0 x} \) and

\[
 f(x|\gamma) = f_0(x) + \Delta \gamma f_1(x) + (\Delta \gamma)^2 f_2(x) + \cdots . \quad (2.17)
\]

If it is justified to neglect the higher powers of \( \Delta \gamma/\gamma_0 \), we can again describe our experimental distribution this time by a superposition of three distributions \( f_0'(x') \), \( f_1'(x') \), \( f_2'(x') \) which are the distorted versions of \( f_0(x), f_1(x), f_2(x) \). The parameter \( \Delta \gamma \) is determined by a \( \chi^2 \) or likelihood fit. In our special case it is even simpler to weight \( f_0 \) by \( x \), and \( x^2 \), respectively, and to superpose the corresponding distributions \( f_0, g_1 = xf_0, g_2 = x^2f_0 \) with the factors given in the following expression:

\[
 f(x|\gamma) \approx f_0(x) \left( 1 + \frac{\Delta \gamma}{\gamma_0} \right) - \gamma_0 g_1(x) \left( \frac{\Delta \gamma}{\gamma_0} + (\frac{\Delta \gamma}{\gamma_0})^2 \right) + \frac{1}{2}g_2(x)\gamma_0^2 \left( \frac{\Delta \gamma}{\gamma_0} \right)^2. \quad (2.18)
\]

The parameter \( \Delta \gamma \) is then modified until the correspondingly weighted sum of the distorted histograms agrees optimally with the data.

Fig. 2.6 illustrates the method with a numerical example. The decays of 5000 events are simulated with a decay rate \( \gamma = 1 \) and a large Gaussian smearing with a standard deviation \( \sigma = 0.5 \). The histogram formed by these events is indicated by the square dots. 100000 Monte Carlo events are generated with a decay rate \( \gamma_0 = 1.2 \) and histogrammed with the weights \( 1, x \), and \( x^2 \). The result of the fit is \( \Delta \gamma = -0.196 \pm 0.020 \) and \( \gamma = 1.004 \pm 0.020 \). The adjusted Monte Carlo simulation is displayed as the black histogram and the contributions by the weighted histograms to the fit are indicated as the red and blue histograms. Due to the large difference between the true value of \( \gamma \) and the one used in the simulation, the quadratic term in \( \Delta \gamma \) has to be included in the expansion. If the Monte Carlo events are generated with \( \gamma_0 = 1.1 \), the linear approximation is sufficient.

### 2.2.2 Data contaminated by background and correlated errors

Often the prediction refers to data that are contaminated by background. If we know the source of the background and its distribution in the true variable \( x \), we can simply reformulate the p.d.f. in such a way that it includes the background. The common case, however, is that the background distribution is known only as a function of the observed variable \( x' \) or has to be estimated from the observed histogram.

We then subtract the estimated background in each bin and include its uncertainty \( \delta(b_i) \) in quadrature in the denominator of (2.7):

\[
\chi^2 = \sum_{i=1}^{B} \frac{(d_i - b_i - c t_i)^2}{c t_i + \delta^2(b_i)}. \quad (2.19)
\]

The normalization constant becomes \( \hat{c} = \Sigma_i(d_i - b_i)/\Sigma_i t_i \). If we have an absolute prediction \( b_i \), for instance from an ancillary measurement with high statistics, the error of Poisson distributed background is \( \delta^2(b_i) = b_i \). If an external background measurement \( b_{0i} \) is Poisson distributed and normalized, \( b_i = \nu b_{0i} \) with given uncertainty \( \delta_\nu \) of the normalization, we get instead of the simple diagonal errors in first order the covariance matrix \( C \):
2.2 Weighting approach

\[
C = \begin{pmatrix}
ct_1 + b_1^2 \delta_1^2 / \nu^2 + (1 + \nu) b_1 & \delta_1^2 b_2 / \nu^2 & \cdots \\
\delta_2 b_2 b_1 / \nu^2 & ct_2 + b_2^2 \delta_2^2 / \nu^2 + (1 + \nu) b_2 & \cdots \\
& \ddots & \ddots \\
& & \ddots & \ddots \\
& & & \ddots & \ddots \\
& & & & \ddots & \ddots \\
& & & & & \ddots & \ddots \\
\end{pmatrix}.
\]  

(2.20)

For a simple numerical example with \( b_i = 20 \), \( \nu = 1 \), \( \delta_{\nu} = 0.2 \) we get \( b_1^2 \delta_1^2 / \nu^2 + (1 + \nu) b_1 = 16 + 40 \) and \( \delta_2 b_2 b_1 / \nu^2 = 16 \).

The modified \( \chi^2 \) expression is:

\[
\chi^2 = \sum_{i=1}^{B} (d_i - b - ct_i) C^{-1}(d_i - b - ct_i)^T.
\]

(2.21)

The inverse \( C^{-1} \) of the covariance or error matrix is the weight matrix.

We have assumed that the errors correspond to Poisson distributions. If this is not true, then (2.20) has to be modified accordingly.

2.2.3 Including the statistical uncertainty of the simulation

One should always attempt to generate as many events as necessary to justify neglecting the statistical uncertainties of the simulation. Usually it is much cheaper to generate a simulated event than an experimental one. However, there are exceptions which we will discuss now. We cannot apply anymore the likelihood method and have to come back to the LS formalism where we have to minimize \( \chi^2 \).

\[
\chi^2 = \sum_{i=1}^{B} (d_i - ct_i)^2 / \delta_i^2.
\]

The denominator \( \delta_i \), the error of \( d_i - ct_i \) now has to include the error of \( t_i \). To evaluate it, we first have to estimate the expected number \( E(d_i) = E(ct_i) = \tau_i \).

Poisson errors

For a few lines we suppress the index \( i \). If not only \( d \) but also \( t \) is a Poisson number, \( t \sim P(t|\tau/c) \) then the result is

\[
\hat{\tau} = \frac{d + t}{1 + 1/c}.
\]

(2.22)

We derive this relation in the Appendix 2. An intuitive justification of (2.22) is the following: If in an experiment \( d \) decays are observed in the unit time interval and \( t \) decays in the interval \((1/c)\), then the rate is \((d + t)/(1 + 1/c)\). We have to add the numbers \( d \) and \( t \) and divide by the total time.

However, the prediction \( t \) is a sum of \( m \) weights, \( t = \sum_j m w_j \), where \( m \) is a Poisson number and also \( w_j \) are i.i.d. random numbers. Thus \( t \) follows a so-called compound Poisson distribution (CPD). The expected value of \( t \) is \( E(t) = E(m)E(w) \) and its variance \( \sigma_{\text{CPD}}^2 \) is equal to \( E(m)E(w^2) \). An obvious estimate of it is the sum of the observed weights squared, \( \sigma_{\text{CPD}}^2 = \sum_{k=1}^{m} w_k^2 = m \bar{w}^2 \). The estimate of the relative error of \( t \) is then

\[
\frac{\delta_t}{t} = \sqrt{\frac{m \bar{w}^2}{m \bar{w}}} = \frac{\sqrt{\bar{w}^2}}{\sqrt{m \bar{w}}}.
\]
The error is larger than the usual Poisson value $1/\sqrt{m}$ by the factor $\sqrt{w^2/m}$. It is convenient to introduce a fictive number of events $\tilde{m}$, the so-called *equivalent number of unweighted events* or *effective number of events*. The number $\tilde{m}$,
\[ \tilde{m} = \frac{m}{\sqrt{\bar{w}^2}}, \]
has the same error $\sqrt{\tilde{m}}$ as a standard Poisson number $\tilde{m}$. Sloppily speaking, the weighted sum $t$ has the same statistical significance as $\tilde{m}$ unweighted events. The effective number of events is of course smaller than the number $m$ of Monte Carlo events. It comes close to it, if the dispersion of the weights is small and it agrees with it, if all weights are equal.

In the large number limit $\tilde{m} \to \infty$ the distribution of $\tilde{m}$ and then of course also that of $t$ follows a normal distribution. In Appendix 1 we show that $\tilde{m}$ can even better be described by a Poisson distribution, and $t$ which is proportional to $\tilde{m}$, $t = s\tilde{m}$, $s = \sqrt{\bar{w}^2}$, by a scaled Poisson distribution (SPD). We can assume that we can always afford a sufficiently large number of Monte Carlo events to justify the asymptotic treatment of the CPD with a scaled Poisson distribution.

With the approximation of the CPD by the SPD we can use the result (2.22) and get for $\tau = E(d) = E(ct) = E(cs\tilde{m})$
\[ \tau = \frac{d + \tilde{m}}{1 + 1/(cs)}. \]
Using $E(d) = \tau$, and $E(\tilde{m}) = \tau/(cs)$ we obtain (see Appendix 1)
\[ \delta^2 = c\left(\sqrt{\frac{\bar{w}^2}{m}}d + t\right). \]
Finally we sum over all bins and get the simple result
\[ \chi^2 = \sum_{i=1}^{B} \frac{(d_i - ct_i)^2}{(c\sqrt{\frac{\bar{w}^2}{m}}d_i + t_i)}. \quad (2.23) \]

**Normal approximation**

For completeness we estimate the uncertainty $\delta^2$ also for the case of normally distributed errors. The normal approximation has to be used if the observed numbers $d_i$ contain correction terms. We omit again the bin index and assume that the expected value $E(d)$ is equal to the expected value $E(ct) = \tau$. For the uncertainty of $d$ we assume that it is proportional to the expected number of $d$, $\delta d^2 = c'\tau$ while for the simulation we stick to the Poisson error $\delta^2_{ct} = c^2 t$. In Appendix 2 we derive
\[ \hat{\tau} = \left[ \frac{cd^2 + c'ct^2}{c + c'} \right]^{1/2} \]
and
\[ \delta^2 = \delta_d^2 + \delta_{ct}^2 = \hat{\tau}(c' + c). \]
In the Poisson limit we get
\[ \hat{\tau} = \left[ \frac{d^2 + ct^2}{1 + 1/c} \right]^{1/2} \]
and
\[ \delta^2 = \delta_d^2 + \delta_{ct}^2 = \hat{\tau}(1 + c). \]

*Remark*: The expected bin content $\tau_i$ in bin $i$ is a nuisance parameter. We have estimated it out, which in principle is a doubtful method. It is justified in our case because the correlation of the parameters $\tau_i$ with $c$ is negligible.
Summary of the procedure

Let us summarize the whole procedure:

1. Simulate the experiment with parameter $\theta_0$ and obtain events $(x_{ik}, x'_{ik})$ were $x'_{ik}$ is the smeared variable of the $k$-th event of the $m_i$ events in bin $i$ of the histogram of $x'$.

2. Select a starting value for the normalization $c_0$.

3. Associate a weight $w_{ik} = 1$ to each event.

4. Compute the mean values $\bar{w}_i, \bar{w}_i^2$ of each bin $i$ and $t_i = \Sigma w_i = m_i \bar{w}_i$.

5. Compute $\chi^2$ according to (2.23).

6. Let Simplex modify $c$ and $\theta$, recompute weights: $w_{ik} = f(x_{ik}|\theta) / f(x_{ik}|\theta_0)$

7. Go to 4. until the minimum of $\chi^2$ is reached.

In most cases we can generate enough Monte Carlo events and apply the simpler error calculation of the previous section.

As an example we choose a superposition of a normal distribution and a uniform background:

Example 7. The 4 free parameters of the following superposition of a normal and a uniform distribution:

$$f(x|\mu, \sigma, \phi) = \phi \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] + (1-\phi); \quad 0 \leq x \leq 1$$

are adjusted. They are the normalization of the observed data to the simulation $c$, the mean value $\mu$ and the standard deviation $\sigma$ of the normal distribution and the fraction $\phi$ of the normally distributed events. With the parameter settings $\mu = 0.5, \sigma = 0.05, \phi = 0.7$ the Gaussian is narrow enough to neglect the tails of the distribution outside the interval $[0, 1]$. The Monte Carlo events are generated with the settings $\mu_{MC} = 1.025\mu, \sigma_{MC} = 1.05\sigma, \phi_{MC} = 1.05\phi$. The parameter $c$ is not interesting and not sizably correlated with the other parameters. Strongly correlated are the estimates of the width $\hat{\sigma}$ and the fraction $\hat{\phi}$. For the selected parameter values the correlation coefficient is of the order of 0.3.

The following table summarizes the results of fits averaged over 100 simulated experiments. The resolutions and the biases of $\mu, \sigma$ and $\phi$ are reported for different combinations of event numbers, $N$ the number of observed events and $M$ the number of Monte Carlo events. The standard deviation of the Gaussian smearing $\sigma_s$, and the number of bins $B$ is given.

| $N$ | $M$ | $B$ | $\sigma_s$ | $\sigma_s$ | $\sigma_s$ | $\sigma_s$ | $\sigma_s$ | $\sigma_s$ |
|-----|-----|-----|-------------|-------------|-------------|-------------|-------------|-------------|
| 100 | ML  | 10000| 0.05 | 0.0142 | 0.0187 | 0.0578 | -0.0001 | 0.0025 | 0.0049 |
| 100 | ML  | 20000| 0.05 | 0.0209 | 0.0149 | 0.0590 | -0.0010 | -0.0008 | -0.0020 |
| 200 | ML  | 10000| 0.05 | 0.0087 | 0.0126 | 0.0436 | 0.0001 | 0.0003 | -0.0008 |
| 200 | LS  | 10000| 0.05 | 0.0090 | 0.0138 | 0.0924 | 0.0003 | 0.0049 | -0.0767 |
| 500 | ML  | 20000| 0.05 | 0.0044 | 0.0066 | 0.0211 | 0.0009 | -0.0006 | 0.0006 |
| 500 | LS  | 20000| 0.05 | 0.0043 | 0.0067 | 0.0491 | 0.0008 | -0.0009 | -0.0064 |
| 1000| ML  | 50000| 0.05 | 0.0037 | 0.0049 | 0.0165 | -0.0002 | 0.0002 | 0.0011 |
| 1000| LS  | 50000| 0.05 | 0.0037 | 0.0049 | 0.0169 | -0.0002 | 0.0000 | -0.0002 |
| 1000| LS  | 10000| 0.05 | 0.0049 | 0.0073 | 0.0243 | 0.0002 | 0.0004 | 0.0029 |
| 10000| LS | 10000| 0.05 | 0.0014 | 0.0020 | 0.0070 | 0.0000 | 0.0002 | 0.0001 |
| 10000| LS | 10000| 0.0  | 0.0013 | 0.0018 | 0.0068 | 0.0001 | 0.0002 | 0.0003 |

For a small number of observed events, enough Monte Carlo events can be generated such that their statistical error can be neglected. This has been done in the examples of the first eight rows.
With only 100 events a LS fit fails, because the number of events per bin is too low. In the LS fit bins with less than 5 events are excluded, which may introduce a bias. The ML fit is always successful. With 200 events the MLE of $\mu$ and $\sigma$ is slightly more precise than the LS fit. The error $\delta_\phi$ is larger in the LSF than in the MLF. The last three rows correspond to situations where the statistical fluctuations of the Monte Carlo events have to be taken into account. Increasing the number of histogram bins slightly reduces the parameter errors if the number of events is large. The numbers of the last row are computed for the limiting case were smearing is absent. As expected the error of the width of the bump is slightly reduced.

As we have 100 simulations per example, the uncertainties of the reported biases are a tenth of the corresponding parameter errors, for example $\delta(b_\mu) = \delta_\mu/10$. Only the result for the fraction parameter $\phi$ in the case of 200 observed events is significantly biased.

Whenever the number of Monte Carlo events is sufficiently high, the ML fit should be preferred to a LS fit. The factor that we need depends on the weight distribution and on how well the parameters used in the simulation agree with the parameter estimates. To check the validity of the approximation, the Monte Carlo sample should be subdivided or even better, bootstrap samples should be fitted [32].

2.3 Summary

We have compared event samples suffering from a limited resolution and from acceptance losses to predictions containing unknown parameters. We distinguish different situations:

1. If the distortions are moderate, a standard $\chi^2$ or likelihood fit can be performed, comparing the prediction to the observed data. Corrections to the result can be derived from a Monte Carlo simulations of the measurement process.

2. If the distribution can be written in the form that the parameter functions factorize,

$$f(x|\theta) = \sum g_i(\theta)f_i(x),$$

the observed distribution can be compared to the superposition $\Sigma_i g_i(\theta)f_i(x')$ where $f_i(x')$ are the folded versions of the functions $f_i$. Histograms $t_i$ corresponding to the functions $f_i(x')$ are generated in a Monte Carlo simulation. The parameter $\theta$ is fitted in a least square or maximum likelihood fit where the observed data $d_i$ are compared to $\Sigma g_i(\theta)t_i$. If the p.d.f. $f(x|\theta)$ is not of the form (2.25), it can be expanded in a Taylor series of $\Delta \theta$ around an estimate $\theta_0$ and then $\Delta \theta$ is fitted. The normalization is a free parameter in the fit.

3. In the standard method, individual Monte Carlo events are weighted with $f(x|\theta)/f(x|\theta_0)$ where $x$ is the undistorted variable and $\theta_0$ the parameter value used in the Monte Carlo simulation. The parameter is adjusted such that the histogram of the weighted events has the same shape as the histogram of the experimental data.

3.1. If the statistical fluctuation of the simulated number of events can be neglected, the parameter is estimated in a Poisson MLF.

3.2. In the rare cases where the statistical uncertainties of the simulation have to be taken into account, the parameter is adjusted in a LSF. The sum of weights in a bin is described by a compound Poisson distribution (CPD) and can be approximated by a scaled Poisson distribution (SPD). The calculation of the denominators (error estimates) of the LS summands based upon a SPD is simpler than with the normal approximation and the result is more precise.
4. Background can be taken into account is a LS fit. The normalization of the background contributions leads to correlations which are included in a weight matrix.

3 Discrete inverse problems and the response matrix

3.1 Introduction and definition

We know turn to the problem of unfolding a distorted distribution for which no parametric prediction for the true distribution is available.

3.1.1 An inverse problem

Folding is described by the integral

\[ g(x') = \int_{-\infty}^{\infty} h(x', x) f(x) dx. \]  

(3.1)

The function \( f(x) \) is folded with a response function \( h(x', x) \), resulting in the smeared function \( g(x') \). We call \( f(x) \) the true distribution and \( g(x') \) the smeared distribution or the observed distribution. The three functions \( g, h, f \) can have discontinuities but of course the integral has to exist. The integral equation (3.1) is called Fredholm equation of the first kind with the kernel \( h(x', x) \). If the function \( h(x', x) \) is a function of the difference \( x' - x \) only (3.1) is denoted as convolution integral, but often the terms convolution and folding are not distinguished. The relation (3.1) describes the direct process of folding. We are interested in the inverse problem: Knowing \( g \) and \( h \) we want to infer \( f(x) \). This inverse problem is classified by the mathematicians as ill posed because it has no unique solution. In the direct process high frequencies are washed out. The damping of strongly oscillating contributions in turn means that in mapping \( g \) to \( f \) high frequencies are amplified, and the higher the frequency, the stronger is the amplification. In fact, in practical applications we do not really know \( g \), the information we have consists only in a sample of observations with the unavoidable statistical fluctuations.\(^1\) The fluctuations of \( g \) correspond to large perturbations of \( f \) and consequently to ambiguities.

The response function often, but not always, describes a simple resolution effect and then it is called point spread function (PSF). There exists also more complex situations like in positron emission tomography (PET) where the relation between the observed distribution of two photons and the interesting distribution of their origin is more involved. In PET and many other applications the variables \( x \) and \( x' \) are multi-dimensional.

3.1.2 The histogram representation

Discretization and the response matrix

The disease of the inverse problem can partially be cured by discretization, which essentially means that we construct a parametric model. We usually replace the continuous functions by histograms.

\(^1\)In the statistical literature the fluctuations are called noise.
which can be written as vectors $\theta$ for the true histogram and $d$ for the observed histogram. The two histograms are connected by the response function, here by a matrix $A$. We get for the direct process:

$$E(d) = A\theta.$$  \hfill (3.2)

$$
\begin{pmatrix}
  d_1 \\
  d_2 \\
  \vdots \\
  d_N
\end{pmatrix} =
\begin{pmatrix}
  A_{11} & \cdots & A_{1M} \\
  A_{21} & \cdots & A_{2M} \\
  \vdots & \ddots & \vdots \\
  A_{N1} & \cdots & A_{NM}
\end{pmatrix}
\begin{pmatrix}
  \theta_1 \\
  \vdots \\
  \theta_M
\end{pmatrix}.
$$

Here $d_i$ is the content of bin $i$ of an observed histogram. $E(d)$ is the expected value. $A$ is called \textit{response or folding matrix} and $\theta_j$ is the content of bin $j$ of the undistorted true histogram that we want to determine.

$$\theta_j = \int_{\text{bin } j} f(x)dx$$

$$E(d_i) = \int_{\text{bin } i} dx' \int_{-\infty}^{\infty} h(x', x)f(x)dx$$

$$A_{ij} = \int_{\text{bin } i} dx' \int_{\text{bin } j} h(x', x)f(x)dx / \theta_j.$$  \hfill (3.3)

The value $A_{ij}$ represents the probability that the detector registers an event in bin $i$ that belongs to the true histogram bin $j$. This interpretation assumes that all elements of $d$, $A$ and $\theta$ are positive. The number of columns $M$ is the number of bins in the true histogram and the number of parameters that have to be determined. The number of rows $N$ is the number of bins in the observed histogram. We do not want to have more parameters than measurements and require $N \geq M$. Normally we constrain the unknown true histogram, requiring $N > M$. With $N$ bins of the observed histogram and $M$ bins of the true histogram we have $N - M$ constraints. The relation between the various histograms is shown in Fig. 3.1.\hfill }
3.1 Introduction and definition

We require that $A$ is rank efficient which means that the rank is equal to the number of columns $M$. Formally, this means that all columns are linearly independent and at least $M$ rows are linearly independent: No two bins of the true histogram should produce observed distributions that are proportional to each other. Unfolding would be ambiguous in this situation but a simple solution is to combine the bins. More complex cases that lead to a rank deficiency never occur in practice. A more serious requirement is the following: By definition of $A$, the observed histogram must not contain events that originate from other sources than the $M$ true bins. In other words, The range of the true histogram has to cover all observed events. This requirement often entails that only a small fraction of the events that contained the border bins of the true histogram are found in the observed histogram. The correspondingly low efficiency leads to large errors of the reconstructed number of events in these bins. Most published simulation studies avoid this complication by restricting the range of the true variable.

Some publications refer to an effective rank and to a null space of the matrix $A$. The null space is spanned by vectors that fulfill $A\theta = 0$. With our definitions and the restrictions that we have imposed, the null space is empty and there is no need to introduce an effective rank.

In particle physics the experimental setups are mostly quite complex and for this reason they are simulated with Monte Carlo programs. To construct the matrix $A$ we generate events following an assumed true distribution $f(x)$ characterized by the true variable $x$ and a corresponding true bin
The detector simulation produces the observed variable $x'$ and the corresponding observed bin $i$. We will assume for the moment that we can generate an infinitely large amount of "Monte Carlo events" such that we do not have to care about statistical fluctuations of the elements of $A$. The statistical fluctuations of the observed event numbers be described by the Poisson distribution.

There is another problem that we neglect but that we have to resume later: The matrix $A$ depends to some extent on the true distribution which is not known in the Monte Carlo simulation. The dependence is small if the bins of the true distribution are narrow enough to neglect the fluctuations of $f(x)$ within a bin. This condition cannot always be maintained.

The need for regularization

The discrete model avoids the ambiguity of the continuous ill-posed problem but especially if the response matrix is large, i.e. the bins are narrow compared to the resolution, the matrix is badly conditioned which means that the inverse or pseudo-inverse of $A$ contains large components. This is illustrated in Fig. 3.2 which shows two different original distributions and the corresponding distributions smeared with a Gaussian $\mathcal{N}(x - x'|0,1)$. In spite of the extremely different original distributions, the smeared distributions of the samples are practically indistinguishable. This demonstrates the sizeable information loss that is caused by the smearing, especially in the case of the distribution with four peaks. Sharp structures are washed out and can hardly be reconstructed. Given the observed histogram with some additional noise, it will be almost impossible to exclude one of the two candidates for the true distribution even with a huge amount of data. Since narrow structures in the true distribution are smeared in the observed distribution and in addition modified by statistical fluctuations, naive unfolding can produce oscillations as shown in Fig. 3.3. Typically, the errors of adjacent bins are strongly negatively correlated. Combining them would reduce the errors considerably and produce a histogram that is closer to the true distribution which is shown as a curve in Fig. 3.3.

If the matrix $A$ is quadratic, we can simply invert (3.2) and get an estimate $\hat{\theta}$ of the true histogram.

$$\hat{\theta} = A^{-1}d.$$  \hspace{1cm} (3.4)

In practice this simple solution usually does not work because, as mentioned, our observations suffer from statistical fluctuations.
3.2 The least square solution and the eigenvalue decomposition

In Fig. 3.4 the result of a simple inversion of the data vector of Fig. 1.1 is depicted. The left-hand plot is realized with 10 bins. It is clear that either fewer bins have to be chosen, see Fig. 3.4 right-hand plot, or some smoothing has to be applied.

### 3.1.3 Expansion of the true distribution

Instead of representing the function \( f \) by a histogram, we can expand it into a sum of functions \( B_i \). The \( B_i \) be normalized, \( \int_{-\infty}^{\infty} B_i(x) \, dx = 1 \).

\[
 f(x) \approx \sum_{j=1}^{M} \beta_j B_j(x) 
\]

(3.5)

The response matrix element \( A_{ij} \) now is the probability to observe an event in bin \( i \) of the observed histogram that originates from the distribution \( \phi_j \):

\[
 A_{ij} = \int_{\text{bin } i} dx' \sum_j \int_{-\infty}^{\infty} h(x', x) B_j(x) \, dx
\]

(3.6)

\[
 t_i \approx A_{ij} \beta_j 
\]

(3.7)

In other words, the observed histogram is approximated by a superposition of the histograms produced by folding the functions \( B_j \). Unfolding means to determine the amplitudes \( \beta_j \) of the basis functions \( B_j \).

In stead of the expansion into orthogonal functions, \( f(x) \) can be approximated by a superposition of basic spline functions (\( b \)-splines). For our applications the \( b \)-splines of order 2 (linear), 3 (quadratic) or 4 (cubic) are appropriate (see Appendix 3).

Unfolding then produces a smooth function which normally is closer to the true distribution than a histogram. The disadvantage of spline approximations compared to the histogram representation is that a quantitative comparison with predictions or the combination of several results is more difficult.

**Remark:** In probability density estimation (PDE) a histogram is considered as a first order spline function. The spline function corresponds to the line that limits the top of the histogram bins. The interpretation of a histogram in experimental sciences is different from that in PDE. Observations are collected in bins and then the content of the bin measures the integral of the function \( g \) over the bin and the bin content of the unfolded histogram is an estimate of the integral of \( f \) over that bin. A function can always be described correctly by a histogram. The description by spline functions is an approximation. This has to be kept in mind when we compare the unfolding result to a prediction.

### 3.2 The least square solution and the eigenvalue decomposition

#### 3.2.1 The least square solution

As mentioned, for a *square matrix* \( A \), \( M = N \) the solution of \( \theta \) is simply obtained by matrix inversion, \( \theta = A^{-1} d \). The error matrix \( C_\theta = A^{-1} C_d (A^{-1})^T \) is derived by error propagation. We omit the calculation. In the limit where there is no smearing, \( A \) is diagonal and describes only acceptance losses.

The choice \( M = N \) is not recommended. For \( M \leq N \) the least square function \( \chi^2_{\text{stat}} \) is given by the following relation:
\[ \chi_{\text{stat}}^2 = \sum_{i=1}^{N} \frac{(t_i - d_i)^2}{d_i} = \sum_{i=1}^{N} \frac{(\sum_{k=1}^{M} A_{ik} \theta_k - d_i)^2}{d_i} . \] (3.8)

If the numbers \( d_i \) are not described by a simple Poisson distribution, we have to insert the weight matrix\(^2\) where \( V = C_d^{-1} \) is the inverse of its error matrix \( C_d \):

\[ \chi_{\text{stat}}^2 = \sum_{i,k=1}^{N} \left[(t_i - d_i)V_{ik}(t_k - d_k)\right] . \] (3.9)

If the data follow a Poisson distribution where the statistics is high enough to approximate it by a normal distribution and where the denominator of (3.8) can be approximated by \( d_i \),

\[ \chi_{\text{stat}}^2 = \sum_{i=1}^{N} \frac{(t_i - d_i)^2}{d_i} = \sum_{i=1}^{N} \frac{(\sum_{k=1}^{M} A_{ik} \theta_k - d_i)^2}{d_i} , \] (3.10)

the least square minimum can be evaluated by a simple linear matrix calculus.

The linear LS solution is given in standard textbooks. We apply the transformations

\[ d \Rightarrow b = A^T V d , \] (3.11)
\[ A \Rightarrow Q = A^T V A . \] (3.12)

We call \( Q \) least square matrix. We get for the expected value of \( b \)

\[ \mathbb{E}(b) = Q \theta \] (3.13)

with the LS solution

\[ \hat{\theta} = Q^{-1} b \] (3.14)

and the error matrix \( C_\theta \) of the solution

\[ C_\theta = Q^{-1} . \]

We have simply replaced \( A \) by \( Q \) and \( d \) by \( b \). Both quantities are then known. The matrix \( Q \) is quadratic and can be inverted if the LS solution exists.

### 3.2.2 Eigenvector decomposition of the least square matrix

To understand better the origin of the fluctuations of the LS solution\(^3\), we factorize the matrix \( Q \) in the following way: The matrix\(^4\) \( Q = U \Lambda U^{-1} \) is composed of the diagonal matrix \( \Lambda \) which contains the eigenvalues of \( Q \) and the matrix \( U \) whose columns consist of the eigenvectors \( u_i \) of \( Q \):

\[ Q = \begin{pmatrix} u_1 & u_2 & \ldots & u_M \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_M \end{pmatrix} \begin{pmatrix} u_1 & u_2 & \ldots & u_M \end{pmatrix}^{-1} . \]

\(^2\)In the literature the error matrix or covariance matrix is frequently denoted by \( V \) and the weight matrix by \( V^{-1} \).

\(^3\)We require that the square \( M \times M \) matrix \( Q \) has \( M \) linearly independent eigenvectors and that all eigenvalues are real and positive. These conditions are satisfied if a LS solution exists.
3.2 The least square solution and the eigenvalue decomposition

The least square solution and the eigenvalue decomposition

\[ Q u_i = \lambda_i u_i = v_i , \ i = 1, \ldots, M . \]  

(3.15)

Software to produce the eigenvector decomposition can be found in most mathematical computer libraries.

In case of eigenvalues that appear more than once, the eigenvectors are not uniquely defined. Linear orthogonal combinations can be created by rotations in the corresponding subspace but they produce the same LS solution.

The solution \( \theta \) can be expanded into the orthogonal unit eigenvectors \( u_i \):

\[ \theta = \sum_{i=1}^{M} a_i u_i , \quad \theta_k = \sum_{i=1}^{M} a_i u_{ik} , \]

\[ a_i = \theta \cdot u_i , \quad a_i = \sum_{k=1}^{M} \theta_k u_{ik} . \]

By construction, the amplitudes \( a_i \) are uncorrelated and the norm \(||\theta||^2 = \sum \theta_i^2\) of the solution is given by

\[ ||\theta||^2 = \sum_{i=1}^{M} a_i^2 . \]

The transformed observed vector \( b \) is
In Fig. 3.5 we present a schematic example of a set of eigenvectors. A contribution $u_i$ to the true histogram as shown on the left-hand side will produce a contribution $v_i$ to the observed histogram. It is of the same shape but reduced by the factor $\lambda_i$ as shown on the right-hand side. The eigenvalues decrease from top to bottom. Strongly oscillating components of the true histogram correspond to small eigenvalues. They are hardly visible in the observed data, and in turn, small contributions $v_i$ to the observed data caused by statistical fluctuations can lead to rather large oscillating contributions $u_i = v_i / \lambda_i$ to the unfolded histogram if the eigenvalues are small. Eigenvector contributions with eigenvalues below a certain value cannot be reconstructed, because they cannot be distinguished from noise in the observed histogram.

The eigenvector decomposition is equivalent to the singular value decomposition (SVD). In the following we will often refer to the term SVD instead of the eigenvector decomposition, because the former is commonly used in the unfolding literature.

**Example 8.** In Fig. 3.6 the 20 eigenvectors of a LS matrix ordered with decreasing eigenvalue are displayed. The response matrix has 20 true and 40 observed bins. The graph is generated from a sample of 100 000 uniformly distributed events in the range of the observed and the true variables $0 < x, x' < 1$. The response function is a Gaussian with standard deviation $\sigma_s = 0.04$. The eigenvectors show an oscillatory behavior where the number of clusters corresponds roughly to the eigenvector index.
Fig. 3.7. Observed eigenvectors 1 (top left) and 20 (top right), eigenvalues (bottom left) and significance of eigenvector amplitudes (bottom right).

In Fig. 3.7 top the eigenvectors 1 and 20 folded with the response matrix are shown. A contribution of eigenvector 20 to the observed histogram is similar to that of noise. The eigenvalues shown at the bottom left graph vary by about three orders in magnitude. This means that a contribution of the eigenvector 20 to the true distribution is suppressed by a factor of 1000 with respect to a contribution of eigenvector 1. The bottom right-hand graph shows the significance of the amplitudes that are attributed to the eigenvectors. Significance is defined as the absolute value of the amplitude divided by its error. As we have indicated above, the significance is expected to decreases with decreasing eigenvalue. Due to the symmetry of the problem, the amplitudes with even index should vanish. Statistical fluctuations in the simulation partially destroy the symmetry. Eigenvector contributions where the significance is below one, are compatible with being absent within one standard deviation.

Example 9. In Fig. 3.8 we compare the eigenvalues, the parameter errors and the significance of the amplitudes of the eigenvectors for the experimental resolution $\sigma_s = 0.04$ (left-hand side) with that of $\sigma_s = 0.08$ (right-hand side). The values are displayed as a function of the eigenvector index for a sample consisting of 5000 events. The eigenvalues decrease by 4.5 decades for $\sigma_s = 0.04$ and by 8.5 decades for $\sigma_s = 0.08$ from the first to the last eigenvector. The errors of the amplitudes of the eigenvectors with large index increase dramatically with $\sigma_s$. The vertical lines in the significance plots indicate the number of eigenvectors that should be retained to obtain the best agreement of the unfolded histogram with the true histogram. (As measure of the quality we use the integrated square error $ISE$ which is explained below.) For $\sigma_s = 0.04$ this are 12 and for $\sigma_s = 0.08$ only 8 eigenvectors. For larger numbers the agreement deteriorates due to spurious oscillations. From the
two plots and the fact that the errors are proportional to the square root of the number of events we can derive that 200 times more events with $\sigma_s = 0.08$ are necessary to obtain results with the same precision as with $\sigma_s = 0.04$.

We conclude that it is difficult to compensate a bad resolution of an experiment by increasing the statistics! We should always make an effort to avoid large smearing effects not only because large event numbers are required but also because the unfolding results then depend strongly on a precise knowledge of the response function.
3.2 The least square solution and the eigenvalue decomposition

3.2.3 The effective number of parameters

When we unfold a histogram, the number of bins of the unfolded histogram is the number of free parameters in the fit. The previous example indicates that the number of parameters that we can determine in a given problem is rather limited. Below a certain eigenvalue $\lambda_k$, all parameters have a significance close to or below one. We define an effective number of parameters $N_{eff} = k$ as the number of parameters with eigenvalues above or equal to this limit but we do not count parameters separated from the dominant parameters by a gap of two or more parameters with significance below one. In this way we exclude parameters with small eigenvalues, even if their significance is above one, because the excess is likely to be caused by statistical fluctuations. For the example of Fig. 3.9 with a uniform distribution the effective number of parameters is $N_{eff} = 17$. Because the significances of parameters 18 and 19 are below one, the parameters with index greater than 19 are not considered even if their significance is greater than one. This definition is to a certain extent arbitrary, but it provides a reasonable estimate of the minimum number parameter that we need to describe the data. There are also parameters left of index 17 that are compatible with being zero. We should not exclude the corresponding contributions, because the reason for the small values of the significance are not large errors, but small values of the fitted amplitudes as is indicated in the right-hand graph. This graph shows that some amplitudes that correspond to small eigenvalues become rather large. This is due to the amplification of high frequency noise in the unfolding. The number of bins in the unfolded distribution should not be much larger than the effective number of parameters, because then we keep too much redundant information, but on the other hand it has to be large enough to represent the highest significant eigenvector. A reasonable choice for the number of bins is about twice $N_{eff}$. The optimal number will also depend on the shape of the distribution.

In the following example we study the dependence of $N_{eff}$ and the significance on the number of true bins.

Example 10. Fig. 3.10 is derived from a superposition of two Gaussians with an exponential distribution, see Fig. 2.4, with $n = 1 000 000$ events distributed with a Gaussian resolution of $\sigma_y = 0.08$ into 100 observed bins. The graph contains results for three different choices of the number of bins of the unfolded histogram, namely 20, 40 and 80. The effective number of parameters is 11 for 20
Fig. 3.10. Significance of the eigenvector contributions for three different binnings (20, 40, 80) of the true distribution. The expanded graph (right hand) indicates that the significance slightly increases with the number of bins.

Fig. 3.11. Significance of the eigenvector contributions for two different resolutions. Reducing the smearing parameter $\sigma_s$ by a factor of two (squares) increases the number of effective parameters from 12 to 17.

and 40 bins. With 80 bins the significance for the dominant eigenvectors is slightly higher than for the lower bin numbers and we get $N_{\text{eff}} = 12$. The parameter $N_{\text{eff}}$ is mainly determined by the experimental resolution and less by the number of observed bins. This is demonstrated in Fig. 3.11. The 80 bin result of the previous figure is compared to the significance obtained with the resolution increased by a factor of two.
3.3 Summary

We have studied the unfolding problem in form of histograms and a response matrix. If the number of events is large enough to apply the normal approximation, it can be solved with a linear least square fit and a simple matrix formalism. The unfolding solution can be expanded in linearly independent vectors, the eigenvectors of the least square matrix or equivalently in those of a singular value decomposition. With decreasing eigenvalues the eigenvectors present more and more oscillations. Due to the experimental smearing, the high frequency components are washed out and difficult to distinguish from noise. As a consequence, only a limited number of eigenvectors can significantly be reconstructed. This number $N_{\text{eff}}$, the effective number of parameters, can be extracted from the significance of the eigenvector coefficients. $N_{\text{eff}}$ depends strongly on the experimental resolution and is independent of the number of bins $M$ used in the unfolded histogram as long as $M$ is larger than about twice $N_{\text{eff}}$. For this reason there is no need to choose the number of bins of the unfolded histogram larger $2N_{\text{eff}}$. The dependence of $N_{\text{eff}}$ on the number of bins $N$ in the observed histogram is negligible if $N$ is greater than about $4N_{\text{eff}}$.

4 Unfolding without explicit regularization

4.1 Introduction

If we want to document the experimental information in such a way that it is conserved for a comparison with a theory that might be developed in future or if we want to compare or combine the data with those of another experiment, the results have to be unbiased. To achieve this condition we could store the distorted data together with the resolution function. Such a procedure is optimal in that no information is wasted but it has severe drawbacks. Two large datasets, the experimental data and the Monte Carlo sample would have to be published and the whole analysis work would be left to the scientist who wants to use the data. A less perfect but simple and more practical way is to unfold the experimental effects and to present the data in form of a histogram together with an error matrix which than can be used in a future analysis. To avoid the unpleasant oscillation that we have discussed in the previous section, we have to choose wide bins. Additional explicit smoothing would bias the data and has to be omitted. We have to accept that some information will be lost.

A third possibility which preserves the information that is necessary for a future quantitative analysis is to unfold the data with a simple explicit smoothing step and to document the smoothing function. A comparison of the experimental result with a prediction is then possible but data of different experiments can not be combined.

In the following we turn to the simple and efficient approach where oscillations are suppressed by using wide bins in the unfolded histogram. We call this procedure implicit regularization. The bin contents are fitted either by minimizing the sum of the least squares or by maximizing the likelihood.

We have discussed the LS method already in the previous chapter. The results of the LS fit are in most cases similar to those of the ML fit. In the asymptotic limit where the event numbers tend to infinity, and where the Poisson distribution can be approximated by a normal distribution, the two methods coincide. Contrary to the LS approach, the MLE usually does not produce negative entries in the unfolded histogram. With the LS method also complex situations can be handled, for instance when background has to be taken care of while the ML method requires Poisson distributed event numbers.
4.2 The maximum likelihood approach

Whenever possible, we should apply a maximum likelihood fit instead of a LS fit. With Poisson distributed event numbers \(d_i\) with expected values \(t_i = \sum_j A_{ij} \theta_j\) the probability to obtain \(d_i\) is

\[
P(d_i) = \frac{e^{-t_i} t_i^{d_i}}{d_i!}
\]

and the corresponding log-likelihood is up to an irrelevant constant

\[
\ln L_{\text{stat}} = \sum_{i=1}^N [d_i \ln t_i - t_i]
\]

\[
= \sum_{i=1}^N \left[ d_i \ln \sum_{j=1}^M A_{ij} \theta_j - \sum_{j=1}^M A_{ij} \theta_j \right].
\]

Maximizing \(\ln L_{\text{stat}}\) we obtain an estimate \(\hat{\theta}\) of the true histogram.

Usually we have of the order of 20 bins and of course the same number of correlated parameters which have to be adjusted. In this situation the fit often does not converge very well. Instead of maximizing the log-likelihood with methods like Simplex, we can compute the solution iteratively.

4.3 The Expectation-Maximization algorithm

The iterations follow the Expectation-Maximization (EM) method \[34\]. The EM algorithm is a general iterative method to maximize the likelihood if there are missing (or latent) variables. It is especially useful in classification problems in conjunction with p.d.f.s of the exponential family.\[1\]. Applied to our unfolding problem, the missing information in step \(k\) is the fractions of events \(p_{ij}^{(k)}\) in an observed bin \(i\) that belong to the true bin \(j\). Hence, there are \(M\) missing variables per bin. Its expected value is \(E(p_{ij}) = A_{ij} \theta_j / \sum_j A_{ij} \theta_j\). The following alternating steps\[2\] are repeated:

- Compute the expected log-likelihood given the actual set of parameters \(\theta^{(k)}\) and the observed data \(d\). The expected number of events that migrates from true bin \(j\) to bin \(i\) is

\[
d_{ij}^{(k)} = d_i p_{ij}^{(k)}
\]

\[
= d_i A_{ij} \theta_j^{(k)}/\sum_j A_{ij} \theta_j^{(k)}.
\]

For a Poisson distribution the log-likelihood to observe \(d_{ij}\) events in bin \(i\) that originate from true bin \(j\) is up to a constant \(\ln L(\theta) = -A_{ij} \theta_j + d_{ij} \ln A_{ij} \theta_j\). The expected log-likelihood of \(\theta\) is

\[\text{To the exponential family belong among others the normal, Poisson, exponential, gamma, chi-squared distribution.}\]

\[\text{This description of the EM algorithm is simplified and adapted to our specific problem.}\]

\[1\]To the exponential family belong among others the normal, Poisson, exponential, gamma, chi-squared distribution.

\[2\]This description of the EM algorithm is simplified and adapted to our specific problem.
4.3 The Expectation-Maximization algorithm

\[ Q(\theta|\theta^{(k)}) = \sum_{i=1}^{N} \sum_{j=1}^{M} [-A_{ij}\theta_j + d_i^{(k)}] \ln A_{ij}\theta_j \]
\[ = \sum_{i=1}^{N} \sum_{j=1}^{M} [-A_{ij}\theta_j + d_i^{(k)}] \frac{A_{ij}\theta_j^{(k)}}{\sum_{j=1}^{M} A_{ij}\theta_j^{(k)}} \ln A_{ij}\theta_j. \]

- Maximize the expected likelihood \( Q(\theta|\theta^{(k)}) \) and obtain \( \theta^{(k+1)} \). The computation of the maximum of \( Q \) is easy, because the components of the parameter vector \( \theta \) appear in independent summands.

\[ \frac{\partial Q}{\partial \theta_j} = \sum_{i=1}^{N} [-A_{ij} + d_i \frac{A_{ij}\theta_j^{(k)}}{\sum_{m=1}^{M} A_{ij}\theta_j^{(k)}}] = 0, \]
\[ \alpha_j \theta_j^{(k+1)} = \sum_{i=1}^{N} d_i \frac{A_{ij}\theta_j^{(k)}}{\sum_{j} A_{ij}\theta_j^{(k)}}, \]
\[ \theta_j^{(k+1)} = \frac{\sum_{i=1}^{N} A_{ij}\theta_j^{(k)} d_i}{\alpha_j / \sigma_s^2}. \quad (4.2) \]

In the second line we have replaced \( \Sigma_i A_{ij} \) by \( \alpha_j \) the average acceptance of the events of true bin \( j \).

Before the EM method had been invented, the iterative procedure had been introduced independently by Richardson and Lucy [1, 2] specifically for the solution of unfolding problems. Later it was reinvented by Shepp and Vardi [3], Kondor [33], Mülthei and Schorr [5] and D’Agostini [12] made it popular in particle physics. That the result of the iteration converges to the maximum likelihood solution, is a general property of the EM method but was also proven by Vardi et al. [4] and later independently by Mülthei and Schorr [5]. For a discussion of the application to unfolding see [35].

In the particle physics community, unfolding with the EM method is also called D’Agostini unfolding, Bayesian unfolding, iterative unfolding or Richardson-Lucy unfolding [3, 4]. I propose to agree on the term EM unfolding.

The iterative method to find the MLE is not only extremely simple, it is also fast. Approximately \( 10^4 \) iterations are executed per second on a simple laptop computer. The convergence can be accelerated by choosing a starting distribution that is close to the expected true distribution.

Figure 4.1 illustrates the convergence of the iteration sequence for an example generated according to the distribution of Fig. 4.2. The resolution was set to \( \sigma_s = 0.8 \). The observed and the true distributions had 40 and 18 bins, respectively. A uniform starting distribution is used. The required number of iteration steps increases with the number of events and the smearing parameter \( \sigma_s \).

Example 11. We simulate data according to the distribution
\[ f(x) = 0.5 \left[ 1 + \frac{1}{\sqrt{2\pi}0.08} \exp \left( -\frac{(x - 0.5)^2}{2 \cdot 0.08^2} \right) \right]. \]

\[ ^3 \]In most of the the figures of this report the abbreviation R-L is used.
which is a superposition of a uniform and a normal distribution in the interval $[0, 1]$. It is displayed in Fig. 4.2. The tails of the normal distribution outside the interval are neglected in the normalization. The response function is also a normal distribution with standard deviation $\sigma_s = 0.04$. A total of 100000 events is generated. The observed smeared distribution with 40 bins and the unfolded distribution with 18 bins are shown in Fig. 4.3. The true distribution, displayed in Fig. 4.2, is not much modified by the smearing. The height of the peak is slightly reduced, the peak is a bit wider and at the borders there are acceptance losses. The central plot shows the unfolded distribution with the diagonal errors. Due to the strong correlation between neighboring bins, the errors are about a factor of five larger than $\sqrt{\theta_i}$. In the right-hand plot the correlation coefficients of bin 10 relative to the other bins are given. The correlation with the two adjacent bins is negative. It oscillates with the distance to the considered bins. The correlation coefficients depend only on the bin width and the smearing function and are independent of the shape of the distribution.
4.3 The Expectation-Maximization algorithm

Fig. 4.2. Superposition of a uniform and a normal distribution. The dashed and dotted curves correspond to the solid curve smeared with Gaussian resolutions $\sigma_s = 0.04$ and $\sigma_s = 0.08$, respectively.

Fig. 4.3. Unfolding without explicit regularization. The left-hand plot shows the observed distribution, the central plot is the result of the unfolding for $\sigma_s = 0.04$ and the right-hand plot indicates the correlation of bin 10 with the other bins of the histogram. The curve represents the true distribution.

In the following examples we derive the parameters of the true distribution from the unfolded histogram.

**Example 12.** We reduce the number of events to 5000 and increase the smearing to $\sigma_s = 0.8$ equal to the standard deviation of the peak. To avoid oscillations in the unfolded distribution, the number of true bins is reduced to 8. The location of the peak is shifted to $\mu = 0.53$. The observed histogram, the unfolded histogram and the correlation with respect to bin 4 are presented in Fig. 4.4. The experiment is simulated 500 times and the unfolding result is then used to fit the parameters of the true distribution. The fit results are summarized in Fig. 4.5. The four parameters of the distribution, e.g. the location and width of the peak and the numbers of events in the uniform and the Gaussian part of the distribution, are well reproduced. The biases are negligible compared to the uncertainties. The standard deviations of the distributions of the fitted parameters agree with the error estimates.
Fig. 4.4. Unfolding without explicit regularization. The left-hand plot shows the observed histogram for $\sigma_s = 0.08$, the central plot is the unfolded histogram and the right hand plot indicates the correlation of bin 4 with the other bins of the histogram. The curve represents the true distribution.

Fig. 4.5. Distribution of the fitted parameters of the one-peak example from 500 experiments with 5000 events, 8 true bins and resolution $\sigma_s = 0.08$. The mean value and the standard deviation of the fitted parameters are given in the graphs.
4.3 The Expectation-Maximization algorithm

Fig. 4.6. Two-peak distribution. The solid curve corresponds to the true distribution, the dashed curve contains the experimental smearing.

from the individual fits. The low number of bins causes some loss in resolution. In the Appendix the dependence of the resolution of the location and the width of a Gaussian peak is estimated. In our case the ratio of bin width $w_b$ and observed width of the peak is $w_b/\sqrt{\sigma^2 + \sigma_s^2} = 1.105$. The reductions in resolution by about 5% and 10% for the peak location and the width compared to the ideal situation are moderate.

Example 13. We borrow a more involved example from [22]. The distribution now contains a superposition of a uniform distribution and two normal distributions,

$$f(x) = c_1 \mathcal{N}(x; \mu_1, \sigma_1) + c_2 \mathcal{N}(x; \mu_2, \sigma_2) + c_3 \mathcal{U}(x),$$

with 7 parameters. We keep 40 bins for the observed histogram and choose 10 bins for the unfolded histogram. 5000 events are generated in the range $-7 \leq x \leq 7$ according to $f(x)$ with the parameters $c_1 = 0.2$, $c_2 = 0.5$, $c_3 = 0.3$, $\mu_1 = -2$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 1$. The smearing resolution $\sigma_s = 1$ is chosen equal to the standard deviations of the two peaks. The experiment is repeated 500 times and each time the parameters are estimated. In 10 experiments or 2% of the cases the standard ML fit failed, mainly because the left-hand peak was not well separated from the larger right-hand peak. The results of the successful experiments are summarized in Fig. 4.7. The observed bias is again small compared to the statistical error.

The examples demonstrate that a relatively small number of bins in the true histogram is sufficient to infer several parameters of a theoretical prediction. The precision of the estimates is close to the limit imposed by the statistical fluctuations of the data. The bias of the results is small compared to the statistical errors.
Fig. 4.7. Fit results for the two-peak example.

Fig. 4.8. Eigenvectors of the unfolded histogram ordered with increasing uncertainty of the coefficients.
4.4 Diagonalizing the error matrix

We have seen that we can obtain an uncorrelated parameter set in the linear LS formalism by the eigenvalue decomposition of the LS matrix $Q$. The inverse $Q^{-1}$ is the error matrix. Small eigenvalues of the error matrix correspond to large eigenvectors of $Q$.

The diagonalization of the error matrix is not restricted to the linear LS formalism. It can also be applied to the ML method with the advantage that small event numbers in the observed histogram can be tolerated. However a sensible result can only be obtained if the unfolded histogram does not contain empty or sparsely populated bins. The solution of the unfolding problem can be formulated as a superposition of the eigenvectors of the error matrix which satisfy $v_i \cdot v_j = \delta_{ij}$. The unfolded distribution $\hat{\theta} = \sum a_i v_i$ is a superposition of the eigenvectors with amplitudes given by $a_i = \hat{\theta} \cdot v_i$.

The amplitudes $a_i$ which replace the parameters $\hat{\theta}_i$ are uncorrelated with errors given by the square root of the eigenvalues of the diagonalized error matrix. Diagonalizing the error matrix, we can again estimate the effective number of parameters $N_{eff}$.

**Example 14.** 100000 events are generated according to $f(x) = 0.5[1 + \mathcal{N}(x|0.5, 0.1)]$, smeared with $\sigma_s = 0.1$ and unfolded. The 18 eigenvectors $v_i$, ordered with increasing eigenvalues, which means with increasing uncertainty of the coefficients, are displayed in Fig. 4.8. With increasing eigenvalue the components of the eigenvectors oscillate more and more. The eigenvalues and significances of the eigenvector amplitudes are displayed in Fig. 4.9. From the plot of the significance we can estimate that this problem has effectively 13 independent parameters.

4.5 Choice of the binning

The essential parameter that we have to fix, is the bin width of the true histogram. The optimal binning depends on the available statistics, the band width of the structure that we want to resolve (which can only be guessed) and the detector resolution.
Fig. 4.10 shows how dramatically the result of the MLF depends on the binning. All plots contain the same data, only the number of bins $M$ increase from 10 to 20 in steps of 2. The correlations and correspondingly the diagonal errors increase. The sequence of the histograms helps to choose a reasonable binning. The diagonal errors should be small enough to justify the application of error propagation in a comparison with predictions. As a rule of thumb, the relative error should be less than 0.5. Sometimes compromises have to be made and larger errors especially in tails of the distribution can be accepted. From the visual inspection of Fig. 4.10 we would probably choose the binning of the fourth or fifth plot.

Another helpful parameter is the effective number of independent parameters $N_{\text{eff}}$ that we have discussed in the previous section and which can be derived from the eigenvalue decomposition of the LS matrix and from the error matrix. The number of bins $M$ has to be larger than $N_{\text{eff}}$ but certainly less than $2N_{\text{eff}}$. The $p$-value of the fit has to be acceptable which puts another lower limit on $M$.

Some physicists prefer to impose a limit on the so-called purity. Loosely speaking, the purity is the fraction of the events that are associated to a certain true bin which actually originate from the corresponding observed bin. For a square matrix $A$ and a bin $i$ it is $A^{-1}_{ii}d_i/\theta_i$. This quantity depends only on the resolution and does not take into account the available statistics. It is of some interest because a low purity signals a strong dependence of the result on a precise knowledge of the response function.

In the following Monte Carlo study the number of bins is varied as a function of the two essential parameters, i.e. the number of events and the resolution.

**Example 15.** We shift the normal distribution $\mathcal{N}(0.53, 0.08)$ of the one-peak distribution in order to avoid specific symmetry preferences. The events are again equally divided into uniformly and normally distributed ones. The results for a resolution $\sigma_r = 0.04$ are shown in Fig. 4.11. The top plots contain 100000 events, the central plots 10000 events and the bottom plots only 1000 events. Within a row, the number of bins decreases from left to right. For each plot the effective number
Fig. 4.11. ML fit results for different bin sizes. The number of events decreases from top to bottom from 100000 to 10000 and 1000. In each plot $N_{\text{eff}}$, the number of constraints in the fit and $\chi^2$ are given.

of parameters $N_{\text{eff}}$, the number of constraints in the fit $N_c$ and $\chi^2$ of the fit are given. $N_c$ is equal to the difference between the number of observed bins and the number of true bins. The observed histogram contains 40 bins and for comparison in some cases 20 bins.

The parameter $N_{\text{eff}}$ cannot be larger than the number of true bins $M$, but otherwise it should be independent of the binning. Due to small statistical fluctuations it varies by one unit. When we look at the first row, we realize that all $\chi^2$ values are acceptable. The chosen number of bins varies between 22 and 18 and is close to the number $N_{\text{eff}} = 17$ of independent parameters. The best choice is close to 19 bins. In the second and the third row with less statistics one would select the second plot with 15 and 12 bins, respectively. The plots 2 and 3 in the second row and the plots 3 and 4 in the third row differ only in the number of bins $N$ in the observed histogram. The two histograms are quite similar indicating that the value of $N$ is of minor importance as long as it is significantly larger than $M$.

In Fig. 4.12 the simulations of the previous figure are repeated with the experimental resolution reduced by a factor of two, $\sigma_r = 0.08$. This leads to a strong reduction of the number of independent parameters $N_{\text{eff}}$.

In the following table we summarizes the results for $N_{\text{eff}}$ for the three different choices of the number of events and the two simulated resolutions.

| number of events | $\sigma_e = 0.04$ | $\sigma_e = 0.08$ |
|-----------------|------------------|------------------|
| $10^3$          | 17               | 9                |
| $10^4$          | 13               | 8                |
| $10^5$          | 10               | 7                |
4.6 Dependence on the Monte Carlo input distribution

In the studies presented so far, the Monte Carlo studies have been performed with the response matrix generated starting from the true distribution. In real experiments it has to be based on some guess of the true distribution and as a consequence the unfolded distribution can be biased.

4.6.1 Uniform Monte Carlo distribution

We now use a uniform distribution to determine the response matrix to study the bias caused by this approximation of the true distribution.

Example 16. We unfold our standard one-peak distribution (2.6) with the parameters, \( \mu = 0.53 \), \( \sigma = 0.08 \) and 100 000 events. (The average bias is independent of the number of events.) We fit the four parameters to the unfolding result: the position of a Gaussian peak \( \mu \), its width \( \sigma \), the number of events \( n_n \) of the normal distribution and the number of uniformly distributed events \( n_u \) and compare them to the nominal values. Due to the approximative symmetry of the problem the quantities \( \mu \), \( n_n \) and \( n_u \) are expected to be unbiased, but \( \sigma \) is biased toward smaller values. The bias is the larger

Fig. 4.12. Same as previous figure but for the smearing constant \( \sigma_s \) increased to 0.08.
Fig. 4.13. Relative bias of the fitted width $\sigma$ of a normal distribution due to a biased response matrix. The upper measurements correspond to a resolution of $\sigma_s = 0.08$, the lower ones to $\sigma_s = 0.04$.

Fig. 4.14. The right-hand plot is obtained by combining bins of the left-hand unfolded histogram. The dots correspond to a simple average, the squares to a weighted average.

the wider the bins are. The relative increase of the width as a function of the number of bins in the interval $0 < x < 1$ is shown in Fig. 4.13 for the two experimental resolutions $\sigma_s = 0.04$ and 0.08. (It is interesting to notice that the fit produces the best results for 30 bins where the distributions show extreme oscillations. Apparently, the error matrix is able to account for the fluctuations even though the diagonal errors are quite large.) It is also clear that the bias can be quite sizable for low event numbers and bad experimental resolutions. For an experiment with 1000 events and $\sigma_s = 0.04$ about 10 bins may be tolerable which leads to a resulting bias of about 15%.

A simple solution of this bias problem is not known to my knowledge. An obvious proposal is to unfold with narrow bins and then to combine bins with the intention to reduce the fluctuations. The content of the bins can either simply be added or a weighted sum can be computed.
Example 17. To test this approach, we choose again our standard example with experimental resolution $\sigma_s = 0.04$, with 60 bins in the observed histogram and 30 bins in the unfolded histogram where always two adjacent bins are combined. The result is then a histogram with 15 bins. Fig. 4.14 shows how extreme the oscillations become with 30 bins. It is astonishing that combining bins leads to qualitatively reasonable distributions. However the computed errors are unrealistically large, probably because simple error propagation fails. A fit of the parameters to these distributions produces strongly biased results for all four parameters with large uncertainties.

To get more insight into the origin of the failure we compare the result of a likelihood fit with 20 bins to a fit with 40 bins where always two bins are combined. The histogram obtained by adding the events from two adjacent bins shown in Fig. 4.15 is much less smooth than the corresponding unfolding result from a direct fit as presented in Fig. 4.11.

We have to conclude that our naive method fails. As long as no satisfactory method is available, we have to model the Monte Carlo input distribution such that it is in agreement with the observed data and conforms to constraints hopefully provided by physics.
4.6.2 Spline approximation of the Monte Carlo input distribution

A uniform distribution is a very bad approximation of most true distributions. A possible way to reduce the bias is to approximate the distribution which is used to determine the response matrix by spline functions. To this end the observed histogram is unfolded with a smoothing algorithm where the unfolded distribution is parametrized by b-splines, see Sect. 3.1.3 and Appendix 3. The result is then used to determine the response matrix with wide bins. The response matrix will still be slightly biased because of the applied smoothing, but the bias of the fitted parameters is small compared to the statistical fluctuations.

Example 18. 1000 events following the one-peak distribution (2.6) are generated with experimental resolutions $\sigma_s = 0.04$ and 0.08. The data are histogrammed into 40 bins. The unfolded distribution is parametrized with 15 quadratic b-splines. The spline coefficients are determined with a MLF with a curvature penalty. The result, shown in Fig. 4.16, is then used to determine the response matrix. Using this matrix a sample of 100000 events is generated and unfolded without explicit regularization with the iterative ML method to determine the remaining bias. From the resulting 10 bin histogram together with the error matrix the 4 parameters of the one-peak example are determined in a LSF. The results are shown in the following table in the last two rows. The width of the peak is still slightly biased, but this bias has to be compared to the statistical fluctuation for 1000 events which would be a factor of 10 larger than the error given in the table. The bias is about a factor 6 smaller than the statistical uncertainty and can be tolerated. Our study corresponds to a small number of events. Of course the spline approximation for the distribution used to compute the response matrix would improve with increased statistics and thus the bias would become smaller.

|            | $\mu$ | $\sigma$ | $\rho$ |
|------------|-------|----------|--------|
| nominal    | 0.53  | 0.08     | 0.5    |
| true spline| 0.5293 (6) | 0.0795 (8) | 0.500 (3) |
| unfold spline $\sigma_s = 0.04$ | 0.5298 (6) | 0.0786 (8) | 0.502 (3) |
| unfold spline $\sigma_s = 0.08$ | 0.5295 (8) | 0.0777 (14) | 0.501 (4) |

Example 19. We repeat the same procedure for an exponential distribution with 40 observed bins. To determine the Monte Carlo input distribution used to compute the response matrix, we generate a data sample of 5000 events drawn from the p.d.f. $5 \cdot \exp(-5x)$, $0 < x < 1$ with resolution $\sigma_s = 0.08$ and unfold it to a superposition of 20 linear b-splines, again with the iterative EM method with early stopping. As always a uniform starting distribution is chosen. The results for 8, 20, 30, 40 and 50 iterations are summarized in Fig. 4.17 and compared to the true distribution. Obviously the response matrix is rather insensitive to the number of applied iterations. The unfolding result for 30 iterations is finally chosen to generate the response matrix. The parameter estimates of the slope parameter averaged over 1000 event samples yielded a bias of 0.005 ± 0.003 to be compared to the 1 st. dev. statistical uncertainty of $\delta = \pm 0.096$ for the fit from a single sample. (With a perfect detector the uncertainty of the slope parameter derived from the unbinned sample would be ±0.072.)

The solution presented here is not very elegant and it is not clear how well it works for more complex distributions. Eventually, systematic errors have to be added to the unfolding result. Most distributions in particle physics will be smoother than the one-peak distribution. Then the proposed method will certainly provide satisfactory results.
4.7 Statistical uncertainties introduced by the simulation

In the majority of all cases it is possible to generate enough simulated events to neglect the statistical fluctuation of the elements of the response matrix. If the number of collected events is very large and comparable to the number of simulated events, it may be necessary to include it. An analytic estimate of the error introduced by the uncertainty of the response matrix is complicated. A simple numeric, but computer time consuming solution of the problem is provided by the bootstrap method [32]:

Let us assume that we have $n$ observed events and $m$ Monte Carlo events used to compute $A$. Then we draw from the set of observed events $n$ events with replacement and from the set of simulation events $m$ events with replacement. This means that some of the drawn events are identical. With this bootstrap set we perform the fit of the unfolded distribution. The procedure is repeated many, say $k = 100$ times. The standard deviation of the distribution of the results for each bin is an estimate of the uncertainty of the true bin content, e.g. of the diagonal error. From the diagonal errors and the correlation matrix which is almost identical in all sets, we can compute the covariance matrix. If the numbers $n$ and $m$ are very large, a faster method is to subdivide the data and the Monte Carlo sample in many, say $k = 100$ subsamples and perform the fit for all subsamples. The fluctuation of the fitted values provides an estimate of the corresponding diagonal errors. These errors then have to be scaled down by a factor $\sqrt{k}$ ($= 10$ if 100 subsamples were used). The covariance matrix is obtained again by scaling the correlation matrix to the diagonal error elements. The relative uncertainty of the error estimate is in both cases $\sqrt{k}$.

4.8 Summary and recommendations

Unfolding with wide bins avoids excessive fluctuation in the unfolded histogram and produces unbiased results with well defined errors. The histogram is obtained with a likelihood fit if the data are
Poisson distributed in the observed histogram or otherwise by a LS fit. The MLE is computed with the iterative the EM procedure. The effective number of parameters \( N_{\text{eff}} \) can be estimated with the eigenvalue decomposition of the LS matrix as was shown in the previous chapter, or from the diagonalization of the error matrix. The number of bins in the unfolded histogram should be larger than \( N_{\text{eff}} \) and small enough to avoid large oscillations and errors that forbid linear error propagation. If the distribution contains narrow peaks, the unfolding result depends on the shape of the distribution that is used to compute the response matrix. This shape can be approximated by spline functions which are fitted to the data in a regularized unfolding procedure. Errors of the unfolded histogram due to statistical uncertainties of the response matrix can be estimated by bootstrap methods.

5 Unfolding with explicit regularization

5.1 General considerations

5.1.1 Regularization methods

The main field where professional unfolding methods are applied lies in image reconstruction. In medicine unblurring of tomographic pictures of arterial stenoses, of tumors or orthopedic objects are important. Other areas of interest are unblurring of images of astronomical surveys, of geographical maps, of tomographic pictures of tools or mechanical components like train wheels to detect damages. Also pattern recognition for example of fingerprints or the iris is an important field of interest. The goal of most applications is to dig out hidden or fuzzy structures from blurred images, to remove noise and to improve the contrast, while in physics applications we are mainly interested in quantitative results. We want to be able to combine data, to estimate parameters and to document the results of experiments. This is achieved with regularization by wide binning as discussed in the previous chapter. On the other hand, we want also to take advantage of the fact that physics distributions are rather smooth. Often we can remove the roughness of an unfolding result without affecting very much the real structures of the true distribution. We then can represent it with many more bins than without a smoothing algorithm and obtain a much clearer picture of it. We have a penalty to pay: We cannot safely quantify the uncertainties of our results. This is a general problem in probability density estimation.

To obtain a smooth distribution, we have to implement a mechanism that suppresses oscillations caused by the noise in the observed data.

In the majority of situations in experimental physics we unfold one-dimensional distributions. Most regularization methods permit to extend their recipes to multi-dimensional problems.

We restrict the study to three different explicit regularization methods in their basic form without sophisticated extensions:

1. **Truncation methods**: In the eigenvalue decomposition of the LS matrix (equivalent to the SVD) low eigenvalue contributions to the unfolding solution are suppressed or eliminated.

2. **Penalty methods**: A penalty term which is sensitive to unwanted fluctuations is introduced in the LS or ML fit of the unfolding solution. Typically, deviations from a uniform or a linear distribution are suppressed. Standard methods penalize curvature, low entropy or a large norm of the unfolding distribution.

3. **Iterative fitting with early stopping**: A smooth distribution is iteratively modified and adjusted to the observation. The iteration process based on the EM method is stopped before ‘unacceptable’ oscillations emerge. Alternatively, the iterative unfolding result is smoothed after each iteration. Then the iteration sequence converges automatically.
Some commercial plotting programs offer smoothing algorithms that can be applied to arbitrary distributions. We will not discuss those because the results are difficult to interpret. A simple bin-by-bin correction method has been used in the past in some particle physics experiments. The ratio of the numbers in the observed and the true histogram in the simulation is used to correct the observed histogram. This approach should be discarded because it can produce wrong or strongly biased results.

In a recent publication [24] external constraints (positivity, monotonicity, convexity) are introduced into the unfolding process of specific distributions like $p_t$ distributions of jets. The method provides global coverage and avoids the regularization with a necessarily doubtfully defined regularization parameter. The user has to be sure that the constraints do not exclude unexpected interesting phenomena.

In the following, we first discuss some general properties of regularization approaches and devote the remaining part to the description of the standard methods and to a comparative study with specific distributions. We will mainly stick to our simple standard example with Poisson distributed data of a distribution consisting of a peak over a uniform background and vary the statistics, the width of the peak and the Gaussian resolution.

5.1.2 Acceptance losses

The correction of acceptance losses is straightforward. The correction can be applied either after unfolding the smearing effects or it can be included in the response matrix. The latter method should be preferred because we want to smooth the true distribution and not the distribution which is affected by the acceptance losses. Furthermore as is emphasized in [20], the acceptance can depend on the observed variable. Then a correction of the true distribution is not sufficient but the losses are correctly taken into account in the response matrix.

5.1.3 Variable dependence and correlations

We have already stressed that smoothness criteria cannot be derived from basic principles. Smoothness is not invariant against variable transformations.

Because of the subjective character of smoothing, we can compare the quality of different methods only with selected examples. However, there is a desirable property of unfolding approaches: They should take the specific properties of the smearing process into account: The events in a bin of the observed histogram originate from different bins of the true histogram. Consequently, in the unfolded histogram the corresponding bin contents are negatively correlated. If the experimental distortions are caused by a simple resolution effect that is described by a point spread function, the correlation occurs predominantly between adjacent bins but in more complicated situations the distortions can lead to correlations between bins that correspond to quite different variable values. In the absence of correlations between bins and especially if the resolution is perfect, smoothing should not be active. This feature is inherent in truncated SVD and the EM unfolding with early stopping, but is not realized in penalty regularization methods.

Most unfolding methods have the convenient property that the unfolding result does not depend on the ordering of the bins in the unfolded distribution. This means that multi-dimensional distributions can be represented by one-dimensional histograms. An exception is regularization with a curvature penalty.

---

1All predictions that disagree with the error bounds in at least a single bin are excluded at the given confidence level.

2Of course this can always be achieved manually by setting the regularization parameter to zero.
The dependence of the smoothness criteria on the chosen variable can be used to adapt the regularization approaches to specific problems. If, for instance, penalties favor a uniform distribution, we can choose a variable in which we expect that the distribution is roughly uniform but in most cases it is better to adapt the penalty function. On might for instance not penalize the deviations from uniformity for a nearly exponential distribution but the deviation from an exponential. The corresponding procedure in the iterative EM method is to select the starting distribution such that it corresponds to our expectation of the true distribution.

5.1.4 Measures of the unfolding quality in Monte Carlo experiments

To get a feeling for a reasonable setting of a regularization parameter, we can compare the unfolded distribution \( \hat{\theta} \) to the true distribution \( \theta \) in toy experiments. For a quantitative comparison, we introduce a variable \( X^2 \), defined by

\[
X^2 = \sum_{i=1}^{N} \frac{(\hat{\theta}_i - \theta_i)^2}{\theta_i}
\]

which resembles the \( \chi^2 \) statistic used in goodness-of-fit tests with Poisson distributed histogram entries. (The choice of the parameter \( X^2 \) as a test quantity is somewhat arbitrary, \( X^2 \) is not \( \chi^2 \) distributed.)

In PDE the standard measure of the agreement between the true and the constructed distributions is the integrated squared error (ISE). For a functions \( f(x) \) and its PDE \( \hat{f}(x) \) it is defined by

\[
ISE = \int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 \, dx .
\]

(Other common measures are the integrated absolute error \( \int |\hat{f}(x) - f(x)| \, dx \) and the Kullback-Liebler distance \( \int [\hat{f}(x) \ln(\hat{f}(x)/f(x))] \, dx \) which is related to the likelihood ratio. Contrary to ISE, these measures are dimensionless.)

The expected value of \( ISE \) is called mean integrated square error (MISE).

\[
MISE = \mathbb{E} \left[ \int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 \, dx \right] .
\]

ISE is not defined for histograms in the way as physicists interpret them. To adapt the ISE concept to our needs, we modify the definition such that is measures the difference of the estimated content of the histogram \( \hat{\theta}_i \) and the prediction \( \theta_i \).

\[
ISE = \sum_{i=1}^{M} (\hat{\theta}_i - \theta_i)^2
\]

In the following sections the value has been normalized to the event numbers and the numbers of bins. In the comparisons only the relative values are important and whether we choose (5.1) or (5.2) which is simpler to compute, does not matter. ISE defined by (5.2) depends on the binning and, as has been pointed out by Volobouev [39], it is zero for a single bin. The ISE attributes more weight to regions where there are many events than \( X^2 \), but the application of the ISE and the quantity \( X^2 \) usually lead to similar conclusions. (The denotation of the modified quantities by the names of well defined parameters ISE and MISE is unfortunate but for technical reasons it cannot be changed any more in this report. The definition (5.2) is sensible if the unfolding result is compared to a prediction but if it is used for the visualization of the distribution or for a simulation the PDE definition (5.1) is relevant.)
5.1.5 Choice of the regularization strength

A critical parameter in all unfolding procedure is the regularization strength which regulates the smoothness of the unfolding result and which determines bias and precision. The optimal value of the regularization parameter depends on the specific application. To fix it, we must have an idea about the shape of the true distribution. We might choose it differently for a structure function, a Drell-Yan distribution with possible spikes, a transverse momentum distribution and the distribution of the cosmic background radiation. We need some prior knowledge. Based on purely statistical arguments, we cannot define smoothness and the optimal regularization parameter. From the data we can only estimate an upper limit of the regularization parameter: The unfolded distribution has to be statistically compatible within the measurements. Most unfolding methods try to approach this limit and eliminate fluctuations that are compatible with noise. There is no scientific justification for this pragmatic choice and one has to be aware of the fact that in this way real structures may be eliminated that can be resolved with higher statistics.

Visual inspection

If we resign to the idea to use the unfolded distribution for parameter fits, it seems tolerable to apply subjective criteria for the choice of the regularization strength. By inspection of the unfolding results obtained with increasing regularization, we are to some extent able to distinguish fluctuations caused by the procedure from structures in the true distribution and to select a reasonable value. Probably this method is in most cases as good as the following approaches which are partially quite involved.

Goodness-of-fit approaches

An obvious quantity to look at is certainly the $\chi^2$ statistic. In a standard LSF, without regularization, with normally distributed errors, $N$ observed quantities and $M$ fitted parameters, we expect that $\chi^2_0$ follows a $\chi^2$ distribution with $NDF = N - M$ degrees of freedom. With regularization the value $\chi^2$ will be larger than $\chi^2_0$ by $\Delta \chi^2$. It is not reasonable to cut on $\chi^2$, the relevant parameter is $\Delta \chi^2$. With the usual approximations, we find 1, 2, 3 standard deviation error limits of the fitted parameters by increasing $\chi^2_0$ by $\Delta \chi^2 = 1, 4, 9$ with confidence $cl$,

$$cl = \int_0^{\Delta \chi^2} f_{NDF}(u) du$$

where $f_{NDF}(u)$ is the $\chi^2$ distribution with $NDF$ degrees of freedom. With $NDF$ of the order of 20, the confidence that the true solution is contained in the one standard deviation interval is small. To be independent of the $NDF$, we turn to the $p$-value defined by

$$p = \int_{\Delta \chi^2}^{\infty} f_{NDF}(u) du = 1 - cl$$

Requiring a certain value $cl$ or $p$, we can derive from (5.4) a corresponding $\chi^2$ boundary $\chi^2_c$ in the $M$-dimensional parameter space. The parameter $p$ corresponds to the frequency in which the true parameter point is located outside the boundary in a large number of measurements.

A LSF or MLF without regularization produces the best estimate of the parameters together with a goodness of fit quantity $\chi^2_0$ and a confidence interval limited by the selected values of $\Delta \chi^2$. A small value of $cl$ and a correspondingly large $p$-value of the regularized parameter indicate that the fit is well compatible with the measurement. We could fix the regularization parameter by choosing
5.1 General considerations

a limiting p-value $p_{\text{reg}}$ which corresponds to a value $\Delta \chi^2_{\text{reg}}$ by which the regularization is allowed to increase $\chi^2$. For example choosing $p_{\text{reg}} = 0.9$, the unfolded distribution should have $\chi^2$ smaller than the true distribution in 90% of all cases.

Unfortunately the parameter estimate $\hat{\theta}$ without regularization is degenerated. LSFs produce bins with negative entries and the related unphysical fit results are hard to accept as best estimates. The MLF avoids negative values, but while the LSF produces negative entries, often empty bins are obtained. The MLF is to be preferred to the LSF but the constraint to positive event numbers entails an increase of $\chi^2$ relative to the nominal value for a LSF. We cannot expect that the expected value of our test quantity $\Delta \chi^2$ obeys a $\chi^2$ distribution and the distribution of the pseudo p-value $p'$ derived from (5.3) will not be uniform. Its distribution can be generated in simulations where we know the true distribution. Since the $p'$ distribution depends mainly on the resolution, the number of events and the binning and less on the shape of the true distribution, it can be estimated also for real experiments.

Example 20. We generate events, perform a MLF and compute $\chi^2_0$. Folding the true distribution and comparing it to the observed distribution we get $\chi^2_{\text{true}}$ which is larger than $\chi^2_0$ by $\Delta \chi^2$. The number $\Delta \chi^2$ is converted to $p'$ by (5.4). The results for the pseudo p-value of the one-peak distribution with 20 true bins, 40 observed bins, different event numbers and resolutions derived from 10000 simulated experiments are displayed in Fig. 5.1. The distributions are far from being uniform. The lower the event number and the worse the resolution is, the more the distribution is peaked towards large pseudo p-values.

It is obvious that we have to take into account the expected shape of the distribution when we derive the regularization parameter. The value of the cut has to be calibrated and it is not guaranteed that the same value is optimal independent of binning, event numbers and resolution.

Regularization based upon $\chi^2$ or a p-value have been proposed in [10, 19].

Truncation approaches

The unfolding result can be expanded into orthogonal components which are statistically independent.

We have studied above the eigenvector decomposition of the modified least square matrix (equivalent to SVD) and realized that the small eigenvalue components $\lambda_i$ cause the unwanted oscillations.
A smooth result is obtained by cutting all contributions with eigenvalues $\lambda_i, i = 1, \ldots, k$) below a cut value $\lambda_k$. This procedure is called truncated singular value decomposition (TSVD). The value is chosen such that eigenvectors are excluded with statistically insignificant amplitudes. The truncation in the framework of the LSF has its equivalence in the ML method. As shown above, we can order the eigenvectors of the covariance matrix of a MLF according to decreasing errors and retain only the dominant components. This method is attractive theoretically, but in concrete applications technical difficulties may arise due to the fact that the error matrix of the unregularized fit may not be well conditioned. A possible way out of the dilemma could be to perform a MLF with a soft regularization by a penalty term and to base the final regularization on the corresponding well conditioned error matrix. The amplitudes corresponding to the orthogonalized covariance matrix which are statistically significant are retained.

The physicist community is still attached to the - for historical reasons - popular linear matrix calculus. Nowadays computers are fast and truncation based on the diagonalized covariance matrix derived from a non-linear LS or a ML fit is probably the better choice than TSVD.

**Empirical Bayes selection**

Kuusela and Paranetos \cite{22} form the product

$$p = p_1(d|\theta)p_2(\theta|r)$$  \hspace{1cm} (5.5)

where the first factor is simply the probability to observe $d$ given the true distribution $\theta$ and corresponds to the usual Poisson likelihood. The second factor is a kind of smoothness probability. It is roughly of the form $\exp[-rR(\theta)]$ with $r$ the regularization parameter and $R$ a measure of the smoothness, here derived from the curvature. The log-likelihood of $p$ corresponds to our relation (5.3). Now the regularization parameter $r$ is chosen such that the probability marginalized with respect to $\theta$ is maximal \cite{3}.

$$\hat{r} = \arg \max_r \int p_1(d|\theta)p_2(\theta|r)d\theta$$  \hspace{1cm} (5.6)

This means that $r$ is derived from the data. The multi-dimensional integral is considered as intractable numerically, but is solved with the EM algorithm. The estimate $\hat{r}$ is plugged into (5.5) to obtain the unfolding result $\hat{\theta}$. The method is attractive because the regularization constant is fixed in a unique way by a simple principle. Naively, one would optimize $r$ directly from (5.5) together with $\theta$ but then the result would be $r = 0$, i.e. no regularization.

To derive error bands with 95% coverage, an iterative bias correction is applied to the point estimate. The correction has a very similar effect as a decrease of the regularization constant. This is shown in the following example.

---

**Example 21.** The distribution \cite{4.3} is used to generate 1000 observed events with a Gaussian resolution of $\sigma = 1$. The events are histogrammed into 40 bins. The unfolded distribution is parametrized by 30 cubic $b$-splines with 28 knots inside the interval $[-7, 7]$ and 2 knots at the borders. Unfolding is performed with a curvature penalty to suppress strong variations of the unfolded distribution. Fig. 5.2 left shows the true distribution and the unfolded distribution together with the results obtained by bias corrections. The bias correction which is computed with $10^6$ simulated events, brings the unfolded distribution closer to the true distribution, but the fake bump at the left-hand side is also enhanced, because it is corrected as well for an assumed bias. A very similar result is obtained by

---

\footnote{The marginalization with respect to $\theta$ protects against overfitting $r$. A justification of the method is indicated in \cite{22} and a relevant reference is given in the publication.}
Fig. 5.2. Bias correction. The left-hand graph shows the true distribution (B), the unfolded distribution (C) and unfolded distributions with sequential bias corrections (D, E, F). The right-hand graph shows the equivalence of the bias correction and a reduced regularization strength.

decreasing the regularization strength as shown in the right-hand graph. A bias corrected distribution is compared to one where only the regularization strength is decreased such that the heights of the larger peaks agree. The two distributions are hardly distinguishable in this specific example.

In this way the final choice of the regularization parameter used to compute location and width of the error bands is shifted to the choice of the number of bias iterations that have to be applied. Thus a part of the beauty of the method is lost. The bias correction is only applied to the error bands, the original point estimate is retained \cite{36}. Thus, the errors are not centered at the point estimates.

**Iterative minimization of the integrated square error ISE**

The best choice of the regularization strength depends mainly on the resolution, i.e. the response matrix and less on the shape of the true distribution. A crude guess of the latter can be used to estimate the regularization parameter $r$. (Here $r$ is a generic name for the number of iterations in the EM method, the penalty strength or the cut in truncation approaches.) The true distribution is approximated by the unfolding result and then $r$ can be optimized with simulated experiments. A goodness-of-fit measure $Q$ of the quality of the agreement between the distribution used in the simulation and the distribution obtained in the unfolding has to be defined. A natural choice is $Q = ISE$ as defined above. The process can be iterated, but since the shape of the true distribution is not that critical, this will not be necessary in the majority of cases. The procedure consists of the following steps:

1. Unfold $\mathbf{d}$ with varying regularization strength $r$ and select the “best” value $\tilde{r}$ and $\tilde{\theta}^{(0)}$ by visual inspection of the unfolded histograms.
2. Use $\tilde{\theta}^{(0)}$ as input for typically $n = 100$ simulations of “observed” distributions $\tilde{\mathbf{d}}_i$, $i = 1, n$.
3. Unfold each “observed” distribution with varying $r$ and select the value $\tilde{r}_i$ that corresponds to the smallest value of the goodness-of-fit measure $Q$. The value of $Q$ is computed by comparing the unfolded histogram with $\tilde{\theta}^{(0)}$.
4. Take the mean value $\bar{r}$ of the regularization strengths $\tilde{r}_i$, unfold the experimental distribution and obtain $\tilde{\theta}^{(1)}$. If necessary, go back to 2., replace $\tilde{\theta}^{(0)}$ by $\tilde{\theta}^{(1)}$ and iterate.
The procedure is independent of the regularization method and the quality measure. An example of its application is presented in Sect. 5.2.3.

Dembinski and Roth [20] propose to minimize the $MISE$ in the observed space. This quantity is expected to be strongly correlated to $MISE$ in the true space, however, their starting relation (19) $MISE = \int V \, dx' + ISE$ (in different notation, the first term of the right-hand side is the variance integrated over the observed space) is not correct. If the bias is small compared to the integrated variance, the right-hand side is on average approximately twice as large as the left-hand side.

The $L$-curve approach

Large fluctuation imply a large value of the norm $||\theta^2||_2$ of the solution,

$$||\theta^2|| = \sum_{i=1}^{M} \theta_i^2.$$

For a given number of events $\Sigma_i \theta_i$ in the unfolded histogram, the norm is minimal for a uniform distribution. In most unfolding methods, the norm decreases with increasing regularization. In turn the residual norm

$$||A\theta - d||^2$$

which measures the discrepancy of the solution with the observed histogram increases with the regularization strength. The Tikhonov regularization penalizes solution with a large norm of the solution. A balance has to be found between small $||\theta^2||$ and small $||A\theta - d||^2$. The log-log relation of these two quantities, with the regularization strength as parameter, is called L-curve [27]. Ideally, for a rectangular L-shaped curve one would select the regularization corresponding to the corner of the L-curve where both norms are small. For other regularization penalties the norm $||\theta^2||$ is replaced by the corresponding expression. Physicists would probably replace the residual norm by $\chi^2$.

5.1.6 Error assignment to unfolded distributions

The regularization introduces a bias and decreases the error $\delta_s$ obtained in the fit. The height of peaks is reduced, the width is increased, valleys are partially filled. The true uncertainties $\delta$ depend on the nominal error $\delta_s$ and the bias $b$, $\delta^2 = \delta_s^2 + b^2$. Increasing the regularization strength reduces $\delta_s$ but increases the bias $b$. As the bias cannot be calculated, we have to guess its size when we choose the regularization strength. The art is to find the optimal balance between the nominal error and the assumed bias introduced by the regularization.

Calculation of the nominal error

A scientific measurement is of limited use if it is not accompanied by an uncertainty estimate. We have pointed out above that the assignment of errors to regularized unfolding results is problematic. Putting aside our concerns and neglecting the bias, we can calculate the errors in different ways:

1. The usual way is to apply error propagation starting from the observed data $d$. To be consistent with the point estimate, the best estimate of the folded distribution $\hat{d}_i = \Sigma_k A_{ik} \hat{\theta}_k$ should be used instead. Error propagation is quite sensitive to non-linear relations which occur with low event numbers. To avoid the problem, $\theta$ can be re-fitted starting from $\hat{d}$ and the errors can be provided by the fit program.

Error propagation starting from the observed data insted of the best estimate can lead to inconsistent results. A striking example is known as Peelle’s pertinent puzzle [37].
2. The errors can be derived from the curvature matrix at the LS or ML estimates. This is the standard way in which symmetric error limits are computed in the common fitting programs. In principle also likelihood ratio or \( \chi^2 \) contours can be computed. The parameters \( \theta \) is varied, the corresponding histogram \( d \) is computed and compared to \( \hat{d} = A\hat{\theta} \). Values of \( \theta \) that change the difference \( \ln L(\hat{\theta}) - \ln L(\theta) \) by \( 1/2 \) fix the standard likelihood ratio error bounds.

3. We can use bootstrap resampling techniques \([32]\). In short, the data sample is considered as representative of the true folded distribution. From the \( N \) observed events, \( N \) events are drawn with replacement. They form a bootstrap sample \( d^* \) which is histogrammed and unfolded. This procedure is repeated many times and in this way a set of unfolded distributions is generated. From which the fluctuations, confidence intervals and correlations can be extracted. For example, in a selected bin the standard 68\% confidence interval contains 68\% of the bootstrap results. Alternatively we can start from the fitted unfolded distribution \( dA\hat{\theta} \). Bootstrap histograms are then constructed from a random Poisson process. For bin \( i \) we choose \( d_i^* \sim P_{\hat{d}} \). The latter method includes the fluctuation in the total number of events and is consistent. Bootstrap errors include flip-flop effects: Small differences in the observed distribution can lead to large differences in the eigenvector decomposition and in TSVD to sizable changes in the result. Also a curvature penalty may introduce similar effects.

A more detailed and professional discussion of the error estimation with bootstrap methods is presented in \([22]\).

Contrary to claims \([21]\), the nominal errors of the unfolded distribution can be computed for all unfolding approaches with the three mentioned methods \([3]\). In the EM method, the errors computed with methods 1 and 2 do not include the regularization constraint due to early stopping. Bootstrap sampling takes it into account. The error margins are correspondingly smaller.

Problems related to the errors assignment

Frequentist coverage in the context of unfolding means: For a given coverage probability \( cl \), an arbitrary true distribution \( f(x) \) has to be accepted in the fraction \( cl \) of a large number of experiments, and specifically, parameters of the true distribution estimated from the unfolding result should within the computed error intervals contain the true parameter values in the fraction \( cl \) of many experiments. Coverage is violated on purpose by the regularization, as the goal of regularization is to exclude strongly fluctuating distributions that are compatible with the data. Regularized histograms are biased and correctly calculated errors of biased distributions that do not include a possible bias in the error estimate cannot cover even for smooth distributions, contrary to claims in some publications. That the specific distribution which is selected in a toy Monte Carlo experiments is covered by the unfolding solution within the given errors should be a triviality, but even this is not realized by most approaches.

Another problem is the dependence of the estimated errors on the regularization strength. In Fig. 5.3 the diagonal error of the highest bin of the one-peak example is plotted as a function of the regularization strength, here for entropy regularization. A qualitative estimate of the range of reasonable regularizations is indicated in the figure. The interval corresponds to regularization penalties which lead to a satisfactory agreement of the unfolded distribution with the true distribution. Within this range the error varies considerably. Each unfolding approach has its own error definition, since the regularization excludes variations of the unfolded distribution that are compatible with noise, the smoothing is more effective in low statistics experiments than in experiments with a large number of events. The assigned errors do not scale with the square root of the number of events and thus it is impossible to combine two unfolded data sets, even if they are produced under exactly the

\[5\] In the publications \([12, 17]\) by D’Agostini, the error estimation is not correct.
same conditions. Small real structures of a distribution that are resolved with high statistics may be excluded in experiments with a small number of events.

The errors that we can assign to the unfolded distribution are at most a rough indication of the uncertainties and have to be considered with great care. The presentation of the results in form of histograms indicate only the diagonal errors but the errors of adjacent bins are correlated, in some schemes negatively in others positively which makes a big difference if bins are combined. In publications the covariance matrices or some information about the correlations should be given.

Due to the mentioned correlation problems, sometimes bin-wise coverage is attempted or error bands are introduced [20, 22]. Predictions should then remain for instance in the fraction $\alpha$ of the bins inside the error limits. Error bands with a high confidence limit, say 90%, give a good qualitative indication whether a predictions is incompatible with the unfolding result. However, quantitative conclusions should not be drawn.

Approximate coverage is realized in 5.6 by separating the point estimate from the interval estimate. The regularization for the error estimate is weaker than for the point estimate. In this way the bias introduced in the interval estimate is small. In [24] constraints which may be available in some specific cases are used to guaranty (over)coverage.

The restrictions that we have summarized do not mean that explicit unfolding is obsolete. It permits to discover structures in exploratory experiments and its results help to visualize the unknown distribution much better than the histograms produced by implicit unfolding with wide bins.

5.2 EM unfolding with early stopping

In the previous chapter we have seen that the EM algorithm produces the MLE of the unfolded histogram. To suppress the fluctuations of the MLE we stop the iteration once the result is compatible with the data and the bin-to-bin fluctuations are still acceptable. We have to fix the starting distribution and the stopping condition.
5.2 EM unfolding with early stopping

Fig. 5.4. EM iterative unfolding results for 50000 observations and resolution $\sigma_s = 0.04$. The number of iterations is indicated in the plots.

Fig. 5.5. $X^2$ and the $ISE$ as functions of the number of iteration for 50000 events and resolution $\sigma_s = 0.04$.

5.2.1 A few examples

We resume our standard one-peak example for 50000, 5000 and 500 events and two different Gaussian resolutions of $\sigma = 0.04$ and 0.08. The starting distribution is uniform. The results are summarized in the following figures. The unfolded histograms are compared to the true histogram indicated by squares. The number of iterations is given in each plot. For each of the settings the test quantities as defined above, $X^2$ and the $ISE$ are plotted as a function of the number of applied iterations. The optimal number of iterations is similar for both parameters.
The optimal result for 50000 events and $\sigma_s = 0.04$ is obtained with 3 iterations, but qualitatively with 20 iterations the adjustment is satisfactory, too. If we increase $\sigma_s$ to 0.08 about 30 iterations are required to unfold the data. Generally, the optimal number of iterations increases with the Gaussian smoothing parameter $\sigma_s$ and for $\sigma_s = 0$ no iteration would but necessary. The convergence is always quite fast. With only 500 events, the method does not succeed anymore to reproduce height and width of the peak very well.
5.2.2 Reduced iteration speed

Occasionally, already the first or second iteration minimizes the test quantities. Then it may be desirable to have a finer step size, to slow down the convergence. This is achieved with a modified unfolding function. We just have to introduce a parameter \( \beta > 0 \) into (4.2):

\[
\hat{\theta}_j^{(k+1)} = \left[ \sum_{i=1}^{M} A_{ij} \hat{\theta}_j^{(k)} \frac{d_i}{d_i^{(k)}} / \alpha_j + \beta \hat{\theta}_j^{(k)} \right] / (1 + \beta).
\]
The value $\beta = 0$ corresponds to the original sequence (4.2). The value $\beta = 1$ slows down the convergence by about a factor of two and with $\beta = \infty$ the parameter $\theta$ remains unchanged.

Applied to the one-peak example with 500 events and $\sigma_s = 0.04$ where the minimum of $X^2$ is reached at the second iteration, $\beta = 1$ moves the minimum to the forth iteration. However, the result is slightly worse, the $ISE$ is increased from 0.0487 to 0.0514. The slowing down of the convergence does not simply interpolate between the results of the standard iteration.
5.2 EM unfolding with early stopping

Fig. 5.12. Same as Fig. 5.4 but for 500 events, $\sigma_s = 0.04$.

Fig. 5.13. Same as Fig. 5.5 but for 500 events, $\sigma_s = 0.04$. 

\[ X^2 \] \[ \text{number of iterations} \] 
\[ \text{number of iterations} \]
Fig. 5.14. Same as Fig. 5.4 but for resolution $\sigma_s = 0.08$ and 500 events.

Fig. 5.15. Same as Fig. 5.5 but for resolution $\sigma_s = 0.08$ and 500 events.
5.2.3 Choice of the regularization strength

A heuristic method to estimate the best number of iterations based on the $p'$ dependence on the number of iterations has been proposed in [19]. However, the method does not work, if the starting histogram is close to the true histogram and it has been tested only with very few event samples and distributions. A better way to select the number of iterations is provided by the method which tries to minimize the $ISE$ explained in Sect. 5.1.5.

**Example 22.** To test the method, let us look at the distribution of the lowest $ISE$ that can be obtained with the standard example with 5000 events and smearing resolution $\sigma_s = 0.08$. The results from 10000 simulations are given in Fig. 5.16. Each time the number of iterations $k$ for which the $ISE$ is minimum is selected. The distribution of the minimal values of the $ISE$s (Fig. 5.16 left hand) has a mean value of 0.043. The mean integrated square error ($MISE$) as a function of the number of iterations is shown at the right-hand side of the same figure. The minimum is obtained at 14 iterations. Choosing always $k_0 = 14$ iterations a mean value $MISE = 0.047$ is obtained, not much larger than the mean of the individually optimized $ISE$ values. Apparently, the $MISE$ depends only weakly on the number of iterations. In the range from 10 to 20 iterations it varies by less than 10%.

In real experiments we do not know the true distribution needed to estimate the optimal number of iterations, but a preliminary unfolded histogram $\theta^{(0)}$ can be used instead. Starting from $\theta^{(0)}$ a large sample of observed histograms can be generated from which the number of iterations $k_0$ can be derived that minimizes the $MISE$.

The best estimate of $k_0$ depends little on the number $k$ of iterations used to determine the preliminary histogram $\theta^{(0)}$. This is shown in Fig. 5.17 where $k_0$ is plotted as a function of $k$ for the previous example. The reason behind this behavior is that the optimal regularization depends mainly on the experimental resolution and less on the details of the true distribution. In cases where the dependence is larger, the procedure can be iterated.
5.2.4 Introducing a final smoothing step

It has been proposed [12, 17, 23] to apply after the iteration sequence a final smoothing step: After iteration $i$ the result $\theta^{(i)}$ is folded with a smoothing matrix $g$, yielding $\theta^{(i)'}$, $\theta^{(i)'} = \sum_k g_{kl} \theta^{(i)}_l$. If $\theta^{(i)'}$ agrees with $\theta^{(i-1)'}$ within given limits, the iteration sequence is terminated. In this way, convergence to a smooth result is imposed. In [17] it is proposed to add after the convergence one further iteration to $\theta^{(i+1)'}$.

The parameters of the smoothing matrix which define the regularization strength have to be adjusted to the specific properties of the problem that has to be solved. The approach may be very successful in problems where prior knowledge about the shape of the true distribution is available, but in the general case it is not obvious how to choose the smoothing step. The intention of the additional iteration is to avoid a too strong influence of the smoothing step on the final result [17].

5.2.5 Dependence on the starting distribution

So far we have used a uniform starting distribution for the EM iteration. If there is prior knowledge of the approximative shape of the true distribution, for instance from previous experiments, then the uniform histogram can be replaced by a better estimate. Experience shows that the influence of the starting histogram on the unfolding result is rather weak.

Example 23. We repeat the unfolding of the distribution with 50000 events and experimental resolution $\sigma_s = 0.08$ starting with the histogram of the observed events. (The choice of the example with a large number of events is less sensitive to statistical fluctuations than an example with low statistics and should indicate possible systematic effects.) The two results displayed in Fig. 5.18 are qualitatively indistinguishable. Starting with the uniform histogram, the lowest $ISE = 0.0964$ is obtained after 40 iterations with $\chi^2 = 35.1$. With the observed histogram the values obtained after 38 iterations are $ISE = 0.0940$ with the same value $\chi^2 = 35.1$. In the low statistics example with 500 events and resolution $\sigma_s = 0.04$ the minimum is reached already after 2 iterations with the $ISE = 0.0488$ and 0.0487, respectively and values $\chi^2 = 36.0$ and 36.3.
5.3 SVD based methods

SVD unfolding was first applied in particle physics by Hoecker and Kartvelishvili.

5.3.1 Truncated SVD

The SVD decomposes the unfolded histogram into statistically independent vectors, \( \theta_0 = \sum_{i=1}^{M} a_i u_i \), and provides an ordering of the vectors according to their sensitivity to noise. In this way it offers the possibility to obtain a stable solution by chopping off eigenvectors with low eigenvalues. Only contribution with eigenvector indices less than or equal to the index \( m \) are kept:

\[
\theta_{reg} = \sum_{i=1}^{m} a_i u_i .
\]

The choice of the cut-off \( m \) is based on the significance \( S_i = a_i / \delta_i \) of the eigenvector contributions \( a_i \) which is provided by the LS fit. The amplitudes of the eliminated eigenvectors should be compatible with zero within one or two standard deviations.

The application of the method, called truncated SVD (TSVD) is simple and computationally fast. The idea behind TSVD is attractive but it has some limitations:

- The SVD solution is obtained by a linear LS fit. This implies that low event numbers in the observed histogram are not treated correctly. Combining bins with low event numbers can reduce the problem.
- The eigenvalue decomposition is mainly related to the properties of the response matrix and does not sufficiently take into account the shape of the unfolded distribution. Small eigenvalues may correspond to significant structures in the true distribution and the corresponding eigenvectors may be eliminated by the truncation. The combination of the vectors belonging to several "insignificant" amplitudes may contribute significantly to the true distribution.

Figs. 5.19 to 5.29 show unfolding results from TSVD. The same data as in the previous section have been taken. In all figures the central histogram corresponds to the smallest value of the ISE.
Unfolding with explicit regularization

Fig. 5.19. Truncated SVD unfolding results for 50000 events and resolution $\sigma_s = 0.04$. The number of eigenvectors that have been included is indicated. The central plot corresponds to the minimum of the $ISE$.

Fig. 5.20. Distribution of $X^2$ and $ISE$ as a function of the number of the included eigenvectors for 50000 events and resolution $\sigma_s = 0.04$.

Fig. 5.21. Same as Fig. 5.19 but with resolution $\sigma_s = 0.08$ and 50000 events.
5.3 SVD based methods

Fig. 5.22. Same as Fig. 5.20 but with resolution $\sigma_s = 0.08$ and 50000 events.

Fig. 5.23. Same as Fig. 5.19 but with 5000 events and resolution $\sigma_s = 0.04$.

Fig. 5.24. Same as Fig. 5.20 but with 5000 events and resolution $\sigma_s = 0.04$. 
Fig. 5.25. Same as §5.19 but with 5000 events and resolution $\sigma_s = 0.08$.

Fig. 5.26. Same as Fig. 5.20 but with 500 and resolution $\sigma_s = 0.08$.

Fig. 5.27. Same as §5.19 but with 500 events and resolution $\sigma_s = 0.04$. 
5.3 SVD based methods

Fig. 5.28. Same as 5.20 but with 500 events and resolution $\sigma_s = 0.04$.

Fig. 5.29. Same as 5.19 but with 500 events and resolution $\sigma_s = 0.08$.

Fig. 5.30. Same as 5.20 but with 500 events and resolution $\sigma_s = 0.08$. 
that is attainable. The numbers in the plots indicate how many eigenvectors have been included. The results for the examples with only 500 events are unsatisfactory, probably because in the uniform part of the histogram only 12.5 events per bin are expected which makes a linear LSF problematic. In the two examples the reconstructed number of events is significantly higher than the true number. The increases are 5% and 7% for the resolutions \( \sigma_s = 0.08 \) and \( \sigma_s = 0.04 \). The agreement of the unfolded histogram with the true histogram is worse than in the EM method. The values of \( X^2 \) and ISE as functions of the number of eigenvectors are not very smooth and make it difficult to choose the number of retained eigenvectors.

### 5.3.2 Smooth truncation

It has been proposed \[27,38\] to replace the brut force chopping off of the noise dominated components by a smooth cut. This is accomplished by filter factors

\[ \varphi(\lambda) = \frac{\lambda^2}{\lambda^2 + \lambda_0^2} \]  

(5.7)

where \( \lambda_0 \) is the eigenvalue which fixes the degree of smoothing and \( \lambda \) is the eigenvalue corresponding to the coefficient which is to be multiplied by \( \varphi(\lambda) \). The solution is then

\[ \theta_{\text{reg}} = \sum_{i=1}^{M} \varphi(\lambda_i) a_i u_i . \]

The function 5.7 is displayed in Fig. 5.31. The amplitude of the eigenvector with eigenvalue \( \lambda = \lambda_0 \) is reduced by a factor 2. For large eigenvalues \( \lambda \) the filter factor is close to one and for small values it is close to zero. The SVD components with large eigenvalues are hardly affected while components with small eigenvalues are strongly damped. As an example we simulate and unfold 5000 events with resolution \( \sigma_s = 0.04 \) and 50000 events with resolution \( \sigma_s = 0.08 \). As before, the minimum of the ISE is determined by varying the parameter \( \lambda_0 \), see Figs. 5.33 and 5.34. The central plots of Figs. 5.32 and 5.34 correspond to the minima of the ISE. The locations of the minima of \( X^2 \) and ISE differ considerably. The ISEs in this specific example are 0.224 and 0.639, considerably larger than those of the standard truncated SVD which are 0.098 and 0.376. This strong effect may be accidental.

It is a strange compromise to reduce the amplitude of a component \( m \) and to include a fraction of the amplitude of a less significant component \( n > m \).

In [27] it is shown that the filtered SVD solution is equivalent to Tikhonov’s norm regularization under the condition that the uncertainties of the observations correspond to white noise (normally
5.3 SVD based methods

Fig. 5.32. SVD unfolding with smooth truncation. The data sample contains 5000 events, the resolution is $\sigma_s = 0.04$.

Fig. 5.33. SVD unfolding with smooth truncation. $X^2$ and ISE for 5000 events and resolution $\sigma_e = 0.04$.

Fig. 5.34. SVD unfolding with smooth truncation. The data sample consists of 50000 events, the resolution is $\sigma_s = 0.08$. 
5.3.3 Selective SVD

Truncated SVD eliminates eigenvector contributions that suffer from large uncertainties. It happens that modeling the true distribution, high frequency contributions are required and that the corresponding amplitudes are significant. Not the absolute error but the significance is relevant for the decision whether to include a component or not. Instead of a vertical cut in the bottom plots of Fig. 5.3 we could apply a horizontal cut and eliminate for example all components with significance below a certain value, for instance 2 standard deviations (SSVD).

At first sight this idea seems attractive. Why should components that are compatible with noise be kept? Well, typical distributions in particle and astrophysics have no periodic regularities that can be associated to specific frequencies and eigenvectors. If an eigenvector $k_0$ is required to describe the distribution than usually all other eigenvectors $k < k_0$ with lower frequency are required as well. If their significance is low, then also their amplitude is low and they will not strongly influence the result. An exception are some angular distributions with preferred and forbidden frequencies. As a consequence it is reasonable to keep all eigenvectors with eigenvalues larger or equal to the eigenvalue of any significant eigenvector.

In Fig. 5.36, the plot of Fig. 5.3 bottom left is repeated with tenfold statistics. One and two standard deviations are indicated with solid and dashed lines. The dotted line corresponds to $\sqrt{10}$ standard deviations which are expected for a signal of one standard deviation in the corresponding
case with only 5000 events. The observed signals with more than two standard deviations for the eigenvectors 10 and 14 with 5000 events are not present in the high statistics case and therefore obviously due to statistical fluctuations. Independent on whether we choose truncated or selected SVD the scatter plot would suggest to retain 10, 11 or 12 eigenvectors while the minimum of the ISE is obtained with 16 contributions. This shows how difficult it is to select the eigenvector threshold from the significance plots.

Sometimes distributions obey symmetry constraints. For example, a distribution of $x$ may be known to be symmetric with respect to the center $x_c$ of the variable range. Then only odd components should be present and the even eigenvectors can be ignored [27], but a better way to take advantage of the symmetry is to replace $x$ by the absolute value $|x - x_c|$. We can proceed in the same way with more complex symmetries. To infer the symmetry from the observed histogram is dangerous because statistical fluctuations destroy the exact symmetry and then it is impossible to distinguish between fluctuations and real effects.

We conclude that truncated SVD should be preferred to selective SVD.

### 5.4 Penalty regularization

The EM and truncated SVD methods are very intuitive and general. If we have specific ideas about what we consider as smooth, we can penalize deviations from the wanted features by introduction of a penalty term $R$ in the likelihood or LS fit:

$$\ln L = \ln L_{\text{stat}} - R, \quad (5.8)$$

$$\chi^2 = \chi^2_{\text{stat}} + R. \quad (5.9)$$

Here $\ln L_{\text{stat}}$ and $\chi^2_{\text{stat}}$ are the expressions given in (4.1) and (3.8). The sign of $R$ is positive such that with increasing $R$ the unfolded histogram becomes smoother. If we prefer a uniform distribution, $R$ could be chosen proportional to the norm $||\theta||^2 = \sum_{i=1}^{N} \theta_i^2$. This is the simple Tikhonov regularization [7]. Popular are also the entropy regularization which again favors a uniform solution and the curvature regularization which prefers a linear distribution. Entropy regularization is frequently applied in astronomy and was introduced to particle physics by Schmelling. All three methods have the tendency to reduce the height of peaks and to fill up valleys, a common feature of all regularization approaches. More sophisticated penalty functions can be invented if a priori knowledge about the true distribution is available. In particle physics, distributions often have a nearly exponential shape. Then one would select a penalty term which is sensitive to deviations from an exponential distribution.

#### 5.4.1 Curvature regularization

An often applied regularization function $R$ is,

$$R(x) = r_c \left( \frac{d^2 f}{dx^2} \right)^2. \quad (5.10)$$

It increases with the curvature of $f$ and favors a linear unfolded distribution. The regularization constant $r_c$ determines the power of the regularization.

For a histogram of $M$ bins with constant bin width we approximate (5.10) by

$$R = r_c \sum_{i=2}^{M-1} \left( \frac{2\theta_i - \theta_{i-1} - \theta_{i+1}}{n^2} \right)^2. \quad (5.11)$$
Unfolding with explicit regularization

Fig. 5.37. Unfolded histograms with curvature regularization for three different values of the regularization constant. The central plot corresponds to the smallest $ISE$. 50000 events have been generated with resolution $\sigma_s = 0.04$.

Fig. 5.38. Test quantities $X^2$ and the $ISE$ as a function of the regularization constant for 50000 generated events and resolution $\sigma_s = 0.04$.

with $n$ the total number of events and $r_c$ the parameter that fixes the regularization strength.

The Figs. 5.37 to 5.47 show the unfolding results of the same samples as studied in the previous sections. The central histogram corresponds to a minimum of the $ISE$. The regularization parameters used to unfold the histograms are indicated in the figures. (To avoid large numbers, the numbers are not equal but proportional to $r_c$.) Qualitatively the unfolding results are similar to those obtained with the methods discussed above. Below each unfolding histogram $X^2$ and the $ISE$ are presented for varying regularization parameters. The dependence of these quantities on the regularization strength is more complex than in the EM method. The minima of the test quantities are relatively shallow if the event numbers or the resolution are low and occasionally secondary minima occur. The agreement of the unfolded histograms with the true histogram is significantly worse than in the EM approach.

The curvature penalty tries to find a piecewise linear distribution. Sometimes the unfolding result can be disappointing. An example is presented in Fig. 5.49 where a fake triangular peak is generated which is absent in all other smoothing approaches that have been investigated. To illustrate the significance of the excess of events, the nominal errors provided by the fit are included in the graph.
Fig. 5.39. Same as Fig. 5.37 but with resolution $\sigma_s = 0.08$ and 50000 events.

Fig. 5.40. Same as Fig. 5.38 but for resolution $\sigma_s = 0.08$ and 50000 events.

Fig. 5.41. Same as Fig. 5.37 but for 5000 events and $\sigma_s = 0.04$. 
Fig. 5.42. Same as Fig. 5.38 but for 5000 events, $\sigma_s = 0.04$.

Fig. 5.43. Same as Fig. 5.41 but for $\sigma_s = 0.08$ and 5000 events.

Fig. 5.44. Same as 5.42 but for $\sigma_s = 0.08$ and 5000 events.
5.4 Penalty regularization

Fig. 5.45. Same as Fig. 5.37 but with 500 events, $\sigma_s = 0.04$.

Fig. 5.46. Same as Fig. 5.38 but for 500 events, $\sigma_s = 0.04$.

Fig. 5.47. Same as Fig. 5.39 but for 500 events, $\sigma_s = 0.08$. 
5 Unfolding with explicit regularization

Fig. 5.48. Same as Fig. 5.40 but for 500 events, $\sigma_s = 0.08$.

Fig. 5.49. Unfolding with a curvature penalty for two different regularization strength and comparison with entropy regularization.

Special treatment of the border bins

The curvature penalty is a function of the content of three adjacent bins. It is not very efficient at the two border bins of the histogram. In the field of PDE specific methods have been developed to avoid the problem [28]. More smoothing at the edges of the histogram can be achieved by increasing the bin size of the border bins or by increasing the penalty. The latter solution is adopted in [22].

5.4.2 Entropy regularization

We borrough the entropy concept from thermodynamics, where the entropy $S$ measures the randomness of a state and the maximum of $S$ corresponds to the equilibrium state which is the state with the highest probability. It has also been introduced into information theory and into Bayesian statistics to fix prior probabilities. However, there is no intuitive argument why the entropy should be especially suited to cure the fake fluctuations caused by the noise. It is probably the success of the entropy concept in other fields and its relation to probability which have been at the origin of its application in unfolding problems. We penalize a low entropy and thus favor a uniform distribution.
The entropy $S$ of a discrete distribution with probabilities $p_i$, $i = 1, \ldots, M$ is defined through the relation:

$$S = -\sum_{i=1}^{M} p_i \ln p_i .$$

For a random distribution the probability for one of the $n = \sum \theta_i$ events to fall into true bin $i$ is given by $\theta_i/n$. The maximum of the entropy corresponds to an uniform population of the bins, i.e. $\theta_i = \text{const.} = n/M$, and equals $S_{\text{max}} = \frac{1}{M} \ln M$, while its minimum $S_{\text{min}} = 0$ is found for the one-point distribution (all events in the same bin $j$) $\theta_i = n \delta_{i,j}$. We define the entropy regularization penalty with the regularization strength $r_e$ of the distribution by

$$R = r_e \sum_{i=1}^{M} \theta_i \ln \frac{\theta_i}{n} .$$

Adding a term proportional to $R$ to $\chi^2$ or subtracting it from $\ln L$ can be used to smoothen a distribution.

A draw-back of a regularization based on the entropy or the norm is that distant bins are related, while smearing is a local effect. Entropy regularization is popular in astronomy \[6, 16\]. It has been introduced to particle physics applications by Schmelling \[10\].

Simulation results are summarized in the same way as for the curvature regularization in Figs. \[5.50\) to \[5.61\).

### 5.4.3 Tikhonov or norm regularization

The most obvious and simplest way to regularize unfolding results is to penalize a large value of the norm squared $||\theta||^2$ of the solution:

$$R = \frac{r_n}{n^2} \sum_{i=1}^{M} \theta_i^2 .$$

The norm regularization has first been proposed by Tikhonov \[7\]. Minimizing the norm implies a bias towards a small number of events in the unfolded distribution. To avoid this effect, contrary
Fig. 5.51. Test quantities $X^2$ and $ISE$ as a function of the regularization constant for 50000 generated events and resolution $\sigma_s = 0.04$.

Fig. 5.52. Same as Fig. 5.50 but with resolution $\sigma_s = 0.08$ and 50000 events.

Fig. 5.53. Same as Fig. 5.51 but for resolution $\sigma_s = 0.08$ and 50000 events.
Fig. 5.54. Same as Fig. 5.50 but with 5000 events, $\sigma_s = 0.04$.

Fig. 5.55. Same as Fig. 5.51 but for 5000 events, $\sigma_s = 0.04$.

Fig. 5.56. Same as Fig. 5.50 but with 5000 events and resolution $\sigma_s = 0.08$. 
Fig. 5.57. Same as Fig. 5.51 but for 5000 events and resolution $\sigma_s = 0.08$.

Fig. 5.58. Same as Fig. 5.50 but with 500 events, $\sigma_s = 0.04$.

Fig. 5.59. Same as Fig. 5.51 but for 500 events, $\sigma_s = 0.04$. 
5.4 Penalty regularization

Fig. 5.60. Same as Fig. 5.50 but with 500 events and resolution $\sigma_s = 0.08$.

Fig. 5.61. Same as Fig. 5.51 but for 500 events and resolution $\sigma_s = 0.08$.

Fig. 5.62. Norm regularization for 5000 events and resolution $\sigma_e = 0.04$. The central plot correspond to the minimum of the $ISE$. The value of the regularization parameter is indicated in each plot.

to the originally proposed penalty, we normalize the norm to the number of events squared $n^2$. The normalized norm can still bias the result, but in the following examples the bias is negligible.
Fig. 5.63. Distribution of $X^2$ and $ISE$ as a function of the norm regularization parameter with 5000 events and resolution $\sigma_e = 0.04$.

Fig. 5.64. same as Fig. 5.62 but for resolution $\sigma_e = 0.08$ and 5000 events.

Fig. 5.65. same as Fig. 5.63 but for resolution $\sigma_e = 0.08$ and 5000 events.
5.5 Spline approximations

Simulations of particle experiments often are based on PDEs. For instance the proton structure functions are required to predict cross sections in proton proton collisions at the Large Hadron Collider at CERN. For these kind of simulations coarse binned histograms are not optimal and smooth unfolding results are preferred which can be obtained with spline approximations. Unfolding to spline functions was first introduced by Blobel [8]. Spline approximations are also proposed in [22] in conjunction with a curvature penalty and in [20] with entropy regularization. In all three cases the unfolded distribution is approximated by cubic $b$-splines. It is not obvious though that cubic splines are better suited than quadratic splines. The latter are more stable at the border bins of the unfolded histogram. For steeply falling distributions like transverse momentum distributions approximations by linear $b$-splines are appropriate.

The representation of the unfolded function by a superposition of spline functions reduces the dependence of the unfolding result on the function used in the simulation of the response matrix. In the methods with penalty regularization, the construction of a response matrix and the dependence

**Fig. 5.66.** Same as Fig. 5.62 but for 500 events, $\sigma_s = 0.04$.

**Fig. 5.67.** Same as Fig. 5.63 but for 500 events, $\sigma_s = 0.04$.

The unfolding results, especially in the case of 500 events, are less convincing than those of the other penalty methods. We notice a large difference of the penalty values which minimize $X^2$ and $ISE$. 

### 5.5 Spline approximations
of the unfolding result on the distribution used in the Monte Carlo distribution can be avoided altogether with the parameter estimation method explained in Chapter 2. This possibility is also realized in [20] in conjunction with entropy regularization.

It has to be noted that independently of the regularization a systematic error is introduced by the fact that the true distribution is approximated by the spline curve. It can happen that this approximation is poor but normally it is excellent within the statistical uncertainties.

The relevant formulas are given in Sect. 3.1.3.

Once the elements of the response matrix have been computed, the unfolding proceeds in the same way as with histograms. The unfolding procedure is completely analogous to that with histograms. The coefficients \( \beta_i \) are fitted to the observed data vector \( d \) in the same way as in the histogram representation. Alternatively, the coefficients of the b-splines can be directly fitted (see Sect. 2.2).

### 5.5.1 Curvature penalty

The curvature penalty of the spline representation is computed following the analytic method used in [22]. The total squared curvature \( R \)

\[
R = \sum_{ij} \beta_i \Omega_{ij} \beta_j ,
\]

\[
\Omega_{ij} = \int B_i''(x)B_j''(x)dx ,
\]

is a simple function of the second derivatives of the cubic b-splines.

The fit of the b-splines to the data is much less problematic than that of histograms. No secondary minima of ISE are observed.

**Example 24.** We apply the method to three data samples of our standard one-peak example with 5000 events and smearing resolution \( \sigma_s = 0.08 \). Twenty cubic B-splines are fitted to the data. Each time a curvature penalty is applied such that the ISE is minimal.

In Fig. 5.68 the unfolding results are compared to the true distribution. The unfolded distributions are transformed back into the histogram presentation and the ISE values for the histograms are computed. In this way we are able to compare the results to the direct histogram fits. The resulting values are 0.184, 0.134 and 0.121. The ISE values 0.325, 0.279 and 0.445 from the direct histogram fit are in all cases significantly worse than those of the spline fits.

### 5.5.2 EM unfolding

The spline approximation can be implemented in all unfolding methods with minimal changes in the computer programs. Of special interest is the performance of the iterative method which is especially successful in the histogram representation. The likelihood of the spline representation converge to the MLE.

**Example 25.** In Fig. 5.69 results for the EM unfolding are depicted. The same data sets are used as in Fig. 5.68. The convergence of the log-likelihood is displayed in Fig. 5.70. The convergence is initially fast and then the residual value decays exponentially.
A comparison of regularization methods has to be based on selected examples. Examples can be chosen such that a specific penalty function favors the corresponding distribution and thus a comparison is always partially biased. Furthermore, a quantitative comparison of the different unfolding methods is difficult, because often clear rules how to choose the regularization strengths are missing. To avoid this problem, we always select the regularization such that the $ISE$ is minimal. In the following we compare in the Monte Carlo simulations always the best (relative to the $ISE$) achievable unfolding results. To judge the quality, we compare the $ISE$, $X^2$ and $\chi^2$ which is computed by folding the solution and comparing it to the observed histogram. In all cases the reference distribution is assumed to be uniform. The observed and the true histogram contain always 40 and 20
bins, respectively. While the value of the $ISE$ depends on the binning, the quality comparison of the methods is expected to be rather insensitive to it.

We restrict ourselves to the basic versions of the EM method, the simple TSVD and to ML fits with curvature, entropy and norm penalty terms. The adjustment of the regularization parameter to a minimal $ISE$ requires many fits. For this reason, we limit the comparison to only a few examples.
5.6 Quality tests of the unfolding methods and comparison

### Table 5.1. One-peak example, \(X^2\), \(ISE\) and \(\chi^2\), resolution \(\sigma_s = 0.04\)

| statistic | \(\chi^2\) | \(ISE\) | \(\chi^2\) | \(MISE\) | \(\chi^2\) |
|-----------|-------------|----------|-------------|-----------|-------------|
| EM        | 15.1        | 0.049    | 36          | 24.3      | 0.049       | 36          | 24.6      | 0.067     | 24          |
| TSVD      | 25.0        | 0.078    | 34          | 35.2      | 0.098       | 37          | 71.2      | 0.183     | 21          |
| curvature | 14.2        | 0.044    | 42          | 36.2      | 0.083       | 35          | 51.4      | 0.202     | 24          |
| entropy   | 34.0        | 0.122    | 50          | 51.5      | 0.133       | 41          | 91.9      | 0.237     | 29          |
| norm      | 42.4        | 0.123    | 49          | 90.3      | 0.206       | 40          | 116       | 0.262     | 29          |
| observed  | 118         | 0.548    | 483         | 70.9      | 0.191       | 4820        | 454       | 1.230     | -           |

### Table 5.2. One-peak example, \(X^2\), \(ISE\) and \(\chi^2\), resolution \(\sigma_s = 0.08\)

| statistic | \(\chi^2\) | \(ISE\) | \(\chi^2\) | \(MISE\) | \(\chi^2\) |
|-----------|-------------|----------|-------------|-----------|-------------|
| EM        | 23.9        | 0.098    | 29          | 73.3      | 0.131       | 47          | 41.2      | 0.096     | 35          |
| TSVD      | 23.1        | 0.090    | 28          | 174.1     | 0.264       | 49          | 253       | 0.376     | 33          |
| curvature | 26.4        | 0.115    | 33          | 117       | 0.199       | 47          | 167       | 0.388     | 33          |
| entropy   | 31.3        | 0.132    | 35          | 104       | 0.173       | 52          | 154       | 0.301     | 37          |
| norm      | 46.5        | 0.174    | 36          | 153       | 0.224       | 48          | 464       | 1.060     | 36          |
| observed  | 138         | 0.558    | 476         | 239       | 0.796       | 4762        | 2629      | 9.08      | -           |

#### 5.6.1 The one-peak distribution

The results for the one-peak example are summarized in Table 5.1 for smearing resolution \(\sigma_s = 0.04\) and in Table 5.2 for \(\sigma_s = 0.08\). The last line shows the result when the observed distribution is taken as true distribution. It is comforting that with the exception of one case, the considered unfolding approaches produces values of the two test quantities that are better than those obtained without unfolding.

The winner of the comparison is the EM method. There is no clear tendency which would allow a classification of the other approaches. A curvature regularization seems to be more successful than the norm regularization. It is astonishing that the performance of truncated SVD is relatively good for the cases with only 500 events where we would expect that the linear least square fit is inferior to the methods based on the likelihood function.

Another quantity of interest is the unfolding bias with respect to the total number of events. Truncated SVD regularization in the 500 event sample looses 7\% of the events for resolution \(\sigma_s = 0.4\) and 5\% for \(\sigma_s = 0.8\). The bias is below 1\% in all other methods.

#### 5.6.2 A two-peak distribution

We turn to the distribution

\[
f(x) = 0.2N(-2, 1) + 0.5N(2, 1) + 0.3U
\]

defined in the interval \([-7, 7]\) with smearing \(\sigma_s = 1\) which has been used in [22]. The function and its smeared version are displayed in Fig. 5.71 top left. The unfolded distributions obtained with the EM, the truncated SVD and three penalty methods for the first of 10 samples with 5000 events are depicted in the same figure. The optical inspection does not reveal large differences between the results. The mean values of of \(X^2\)and \(ISE\) from the 10 samples are presented in Fig. 5.72. They indicates that truncated SVD and curvature regularization perform less well than the other approaches and that the \(MISE\) obtained with the EM method is significantly smaller than the
values of the competing approaches. In Fig. 5.72 the \( MISEs \) of the EM iteration and the entropy penalty fit of 100 samples are compared. There are large fluctuations from sample to sample, the results are correlated but the values of the EM values are always lower than those of the entropy penalty fit method. The goodness-of-fit statistic \( \chi^2 \) evaluated for the minimum of the \( MISE \) is very similar in all approaches.

It has not been attempted to improve also the statistics for the curvature penalty method to a similar extent. This would be very time consuming, because to find the minimum of the \( ISE \)
5.6 Quality tests of the unfolding methods and comparison

Fig. 5.73. Scatter plot of ISE values obtained from 100 simulated experiments for the two-peak example in the EM method and the entropy penalty fit.

Fig. 5.74. Two-peak distribution: Mean values of $X^2$ and ISE from 5 experiments with 500000 events each.

for this method cannot be automatized. Instead we repeat the study with the number of events increased to 500000 and a different weighting of the contributions to the true distribution: $f(x) = 0.3\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1) + 0.2\mathcal{U}$. It is expected that there the fluctuations of the results from sample to sample are smaller.

In Fig. 5.74 $X^2$ and ISE averaged over 5 samples are plotted for the considered regularization methods. Again the results of the EM approach are considerably better than those of all other approaches.

5.6.3 A lifetime distribution

Fig. 5.75 top left shows an exponential distribution $f(x) = e^{-x}$, ranging from zero to infinity. It is observed with a resolution of $\sigma_s = 1$ in the interval $[-1,5]$ which is subdivided into 40 bins. The true histogram contains 20 bins from which 19 are 0.25 units wide. The 20th bin covers all true values from 4.75 to infinity. In this way it is guaranteed that all observed values have a true partner. The last wide bin is excluded from the ISE calculation. 5000 events have been generated.
5 Unfolding with explicit regularization

The unfolding results summarized in Fig. 5.75 demonstrate a rather good performance of the EM approach. The results for the methods with curvature and entropy penalties are similar but slightly worse, while norm regularization is unable to reproduce the true distribution. TSVD fails technically because the linear LS fit cannot cope with the small event numbers in some bins of the observed histogram.

Of course the performance of all approaches could be improved: We could fit the deviation from a first guess of an exponential distribution and penalize the deviations. In real experiments one would anyway fit the relevant parameters with the method described in Chapter 2.

5.6.4 $A_{pt}$ distribution

The following example is similar to the previous one but is of interest because the smearing uncertainty is not constant and because acceptance losses are introduced explicitly.

The distribution of the transverse momentum squared in particle experiments follows in many cases approximately an exponential distribution. In our toy experiment we simulate again a simple exponential $f(x) = \gamma e^{-\gamma x}$ in the interval $[0, \infty]$ with $\gamma = 0.2$. This time we assume that the acceptance $\varepsilon$ is low for small values of $x$, $\varepsilon = 1/[1 + e^{-(x-2)}]$. The sigmoid function produces $\varepsilon(2) = 0.5$, $\varepsilon(-\infty) = 0$ and $\varepsilon(\infty) = 1$. The true distribution $f(x)$, the accepted part and the smeared version are displayed in Fig. 5.76.

Momentum measurements have the tendency that the error increases with the momentum. Here we assume $\sigma_s(x = 1) = 0.8$ and $\sigma_s \sim \sqrt{x}$. For the graphical representation and to simplify the programming, it is useful to have uncertainties which are proportional to the bin size. As in [38] we apply the transformation $y = \sqrt{x}$ to the distribution $f(x)$ and get $g(y) = 2y\gamma e^{-\gamma y^2}$ with smearing uncertainty $\sigma_{sy} = 0.4$ independent of $y$. The unfolding is then performed in $y$. The result can be transformed back to the original variable $x$ where we obtain a histogram with bin sizes that increase.
Fig. 5.76. Transverse momentum distribution. The true distribution corresponds to the full line of the left hand plot. The accepted part is given by the dashed line. Folding with the resolution function produces the dotted line. The unfolded histogram is evaluated in the range limited by the vertical lines. The transformation to the variable $y = \sqrt{x}$ is presented in the right-hand plot.

Fig. 5.77. Unfolded pt distributions. The drop at low pt values is due to the limited acceptance. Top left: EM method, original binning, Top center: EM method with modified binning. The following plots correspond to TSVD, curvature, entropy and norm regularization.
with \( x \). We choose the interval \([2, 6]\) for the observed variable \( y' \) with 40 bins 0.1 wide and discuss the results in the \( y \) system and do not apply the acceptance correction. This is not the way physicists like the presentation, but it better suited for the unfolding study. The pure Poisson fluctuations can be estimated from the event numbers in the linear ordinate and the resolution is constant across the abscissa. The central 18 bins of the 20 bins of the true histogram are 0.2 wide and contained in the range \( 2.2 < y < 5.8 \). The two border bins cover the intervals \([0, 2.2]\) and \([5.8, \infty]\) and are discarded after the fit of the unfolded distribution.

The result of the unfolded \( y \) histogram for the EM method are shown in the top left plot of Fig. 5.77. The reconstruction of the first two bins is not satisfactory. The reason is the low efficiency for the number of events in the first bin. The performance can be improved by subdividing the bin containing the underflow. The procedure has been repeated with 25 true bins and 50 observed bins where now the underflow bin covers \([0, 1.2]\). The remaining bins, starting from 1.2 are 0.2 wide except the last bin which is kept as before. The first 5 bins and the last bin are not considered in the unfolding test.

The central plot of the top row in Fig. 5.77 shows the improved unfolding result of the EM method and the following plots those of TSVD and the three penalty methods (curvature, entropy and norm). The results for the test quantities are displayed in Fig. 5.78. Again EM performs best. Norm regularization is not competitive.

### 5.7 Parameter estimation from unfolded histograms

Cousins has proposed the following bottom-line test [40]: "If the unfolded spectrum and supplied uncertainties are to be useful for evaluating which of two models is favored by the data (and by how much), then the answer should be materially the same as that which is obtained by smearing the two models and comparing directly to data without unfolding ."

In this report a similar test is applied which maintains the idea behind Cousins’ proposal: In parametric models parameter estimates obtained from the unfolded histogram should be as precise as those obtained directly from the data. While Cousins emphasizes goodness-of-fit we concentrate on parameter estimation. The two approaches should lead to the same conclusions.

We use the unfolding results of the five methods as applied to four simulated samples to estimate the 4 parameters of the one-peak distribution. This example is chosen because the smearing and unfolding effects are best studied in examples that have sharp structures.

Simulation results for the single peak example and moderate smearing \( \sigma_s = 0.04 \) are shown in the Figs. 5.79 and 5.80. The errors are derived from the curvature matrix of the log-likelihood.
function at its maximum, except for the bootstrap errors of the EM approach which are obtained from 10000 bootstrap samples derived from the observed sample after each time 4 EM iterations. The correlation of all bins relative to bin 10 indicate how the regularization reduces the negative correlation. As curvature regularization acts locally, the correlation between adjacent bins is positive. The size of the errors in the EM method are not affected by the regularization. The diagonal errors are large and there is a strong negative correlation between adjacent bins.

The results of the parameter fits are displayed in Figs. 5.81, 5.82, 5.83 and 5.84. The fit failed in two cases with the TSVD approach which we do not consider further.

The peak location, Fig. 5.81, is found with similar precision in the EM method and the three penalty methods. The error estimates of the curvature regularization fail to cover the true value which is indicated by the horizontal line. This is no surprise, as the penalty term in the fit reduces the errors. The error assignments obtained of the entropy and norm regularization seem to be adequate.

As expected, the width of the peak, Fig. 5.82, is reconstructed too wide in all methods except in the EM method with bootstrap errors. The small bias in the latter case is hard to explain.

The event numbers are reasonably well estimated in all methods except in TSVD and EM with bootstrap errors.

We conclude that TSVD and EM iterative unfolding with bootstrap errors should be discarded and that the error estimates of the unfolding procedure with a curvature penalty are doubtful. The

---

Fig. 5.79. Unfolded histograms with assigned errors. The squares indicate the true distribution. Below each histogram the correlation of the content of bin 10 with the other bins is plotted.

---

Due to an initial programming mistake, the random data used for the bootstrap approach differ from those used in the other methods.
In the following we compare the parameter estimates for the one-peak example with widths of the peak, \( \sigma = 0.08 \) a) directly from the observed sample with the weighting method, b) from the unfolded histogram with 8 bins without explicit regularization, c) from the unfolded histogram with entropy regularization, including the full error matrix, d) from the unfolded histogram with entropy regularization with the diagonal errors only. 100 experiments with 5000 events are simulated. The results are summarized in Fig. 5.85 and Tables 5.3. The mean values of the fitted location of the peak performance of the remaining methods is similar. There is a sizable and unavoidable bias in the reconstructed width of the peak. The bias is approximately as large as the statistical error. The number of simulated experiments is too small for more detailed conclusions.

Fig. 5.80. Same as previous figure but for penalty regularization.

Fig. 5.81. Fit results for the location of the peak.
and its width are given together with the errors. The corresponding errors for a single experiment are larger by a factor of 10.

The direct fit exhausts all information and is more precise than all other methods. The results obtained for the width with entropy resolution are strongly biased especially if only the diagonal errors are considered. The bias is about as large as the statistical error. We conclude that Cousins’ test fails.

The errors of the fits with implicit regularization by wide binning are moderately larger than those with explicit regularization but are unbiased. They could probably be reduced by optimizing the bin width.
Fig. 5.85. Comparison of fits of the peak location (left) and its width (right) of the one-peak example with $\sigma = 0.08$. From top to bottom: direct fit to the data, implicit regularization by binning, explicit entropy regularization.

| Method          | # experiments | $\mu$  | $\sigma$ | $\delta\mu$ | $\sigma_{\delta\mu}$ | $\delta\sigma$ | $\sigma_{\delta\sigma}$ | bias$_{\sigma}$ |
|-----------------|---------------|--------|----------|--------------|------------------|--------------|-----------------|----------------|
| nominal         |               | 0.5300 | 0.08     |              |                  |              |                 |                |
| direct          | 100           | 0.52997 | 0.00033  | 0.08059      | 0.00052          | 0.00059      |                  |                |
| implicit        | 100           | 0.53070 | 0.00039  | 0.07969      | 0.00077          | -0.00031     |                  |                |
| explicit        | 100           | 0.53052 | 0.00033  | 0.08483      | 0.00052          | 0.00483      |                  |                |
| explicit diagonal| 100          | 0.53097 | 0.00040  | 0.08527      | 0.00063          | 0.00527      |                  |                |

The situation will be different for rather smooth distributions. It can even happen that the regularization penalty reduces the errors to below the statistical limits of an undistorted distribution. This is acceptable for a probability density estimate but not for a standard measurement. Smooth distributions can be unfolded with wide bins without much loss in information and then an explicit regularization is obsolete.

5.8 Summary and recommendations

We have considered five different unfolding methods:
5.8 Summary and recommendations

- **Methods based on Singular Value Decomposition**: The SVD offers the possibility to discard the disturbing high frequency components of the unfolded histogram. Truncated SVD is to be preferred to selective SVD. A smooth cut-off is not recommended. The application of TSVD is technically simple, but from the considered examples we have to conclude that the performance is not competitive with that of the other considered methods. TSVD fails in some cases with low event numbers.

- **EM unfolding**: The EM method is extremely simple, effective and fast. It can be applied also to spline representations of the unfolded distribution. In the considered examples it performs better than the competing methods.

- **Unfolding with curvature regularization**: This popular method favors linear distributions. A positive feature is the local smoothing which is adapted to point spread distortions. It is more difficult to fix the regularization strength than in other methods. Due to the strong bin to bin correlations the fit of the unfolded histogram has to be assisted to find the absolute minimum. Its application to multi-dimensional histograms is problematic. The assigned errors occasionally exclude the true distribution even in toy experiments and in some low statistics examples the method fails completely to reproduce the true distribution. Some of these difficulties are avoided if the distribution is approximated by a superposition of $b$-splines.

- **Unfolding with entropy regularization**: The method, like norm regularization, favors uniform distributions. In most of the considered examples, the unfolding results obtained with entropy regularization are more precise than those with curvature or norm regularization.

- **Tikhonov’s norm regularization**: The method, like norm regularization, favors uniform distributions. In most of the considered examples, the unfolding results obtained with entropy regularization are more precise than those with curvature or norm regularization.

TSVD and the EM method are both based on plausible general concepts while the penalties used to regularize the results, favor specific features of the distributions.

All methods lead to biased results. The nominal errors that can be attributed to the unfolded histogram are to be considered as rough estimates. The size of the error bars depends on the regularization parameter, the errors do not cover and sometimes exclude even rather smooth distributions. The assigned errors are partially subjective and should be considered as an educated guess of the author. An exception are the errors derived in the very special method [24]. In the EM method they cover approximately, but due to the strong negative correlation of adjacent bins, the diagonal errors are large and exaggerate the uncertainties in the graphical representation where only the diagonal errors can be shown. The unfolding results are usually closer to the true distribution than those obtained with the wide bin method. They help to visualize the true distribution, are useful to exclude qualitatively predictions and can be applied to simulate the distribution. The point estimates derived from the unfolded distributions are biased and less precise than the values derived directly from the observed data. The bottom-line test fails. Because of the bias of the point estimates and the partially arbitrary size of the errors, they cannot be used in quantitative goodness-of-fit tests and parameter estimation.

Further requirements and remarks:

- It is recommended to fix the regularization parameter by visual inspection or with the iterative minimization of the $ISE$.

- The unfolding results depend on the distribution used to determine the response matrix. The dependence can be kept small, if in a first step the data are unfolded into a superposition of $b$-splines. The result is then used to determine the response matrix.

- There are indications that a representation of the unfolded distribution in form of a superposition of $b$-splines is closer to the true distribution than that with a histogram. Furthermore in the
methods with penalty regularization it is not necessary to construct a response matrix and the problem related to its Monte Carlo construction is avoided.

- The considered range of the true variable has to cover all observed variable values. In most Monte Carlo studies this problem is avoided by artificially restricting the true space.

- It is important to study possible systematic errors due to the limited knowledge of the response function. The statistical errors introduced by the response matrix can be estimated with resampling techniques.

- Programs off the shelf should be used with caution. The user and not the author of the program has the responsibility of the correctness of the results and therefore must understand the assumptions and approximations that are made.

- It would be interesting to compare the more sophisticated methods [18, 20, 22] to the simple iterative EM unfolding.

- Publications should always include unbiased results like those provided by histograms with wide bins without explicit regularization. In this way the requirements for the presentation of the unfolding results obtained with explicit regularization can be relaxed.

6 Summary of the summaries and a personal recommendation

- If a theoretical prediction exists, parameters can be extracted directly from the distorted data. It is not necessary to unfold and to construct a response matrix.

- The probability density estimates (PDEs) obtained by unfolding smeared data with explicit regularization provide a good illustration of the underlying true distribution but cannot be used to derive quantitative conclusions. The iterative EM unfolding with early stopping performs significantly better than the other considered unfolding methods.

- If no prediction exists, it is recommended to use the EM method with early stopping to obtain a preliminary spline approximation of the true distribution. This result is then used to determine the optimal number of the EM iterations by the iterative optimization of the ISE and to generate an improved response matrix. The data are unfolded a second time, again to a spline approximation. Qualitative error bands derived with bootstrap resampling methods can be associated to the result. They should be accompanied by an explanation of their relevance. They can be obtained with bootstrap resampling. From the regularized unfolding result a final response matrix is constructed which then is used to unfold the data into an effectively unbiased histogram with wide bins and to compute an error matrix. In this way the probability density estimate (PDE) is accompanied by the information that is needed to estimate parameters and to perform goodness-of-fit tests.

It is hoped that this report will stimulate further systematic studies.

Acknowledgement

I am very grateful to Mikael Kuusela and Victor Panaretos for their patience in explaining to me their unfolding methods and for related discussions. I thank Hans Dembinski for an interesting exchange of ideas concerning the optimization of the regularization, Bob Cousins for helpful comments and his interest in my work, Gerhard Bohm and Vato Kartvelishvily for useful comments. I am especially
grateful to Igor Volobouev for pointing out to me some mistakes in formulas of this report and I appreciate his critical comments.

Appendices

Appendix 1: The Compound Poisson Distribution

If a parameter of a distribution is itself randomly distributed, then we have a compound distribution. The \textit{compound Poisson distribution} (CPD) describes the sum of a Poisson distributed number of very independent and identical distributed weights. In the case we are interested in, we have a continuous weight distribution. A number \( n \) of weights \( w_i, i = 1, \ldots, n \), are randomly chosen from a p.d.f. \( g(w) \) where \( n \) follows a Poisson distribution \( P(n|\lambda) = e^{-\lambda} \lambda^n/n! \). The distribution \( f(x) \) where \( x = \sum_{i=1}^{n} w_i \) has a complicated analytic form, but its moments are easy to calculate.

The first two moments of \( f(x) \) follow directly from the definition:
\[
\mu = \lambda E(w), \\
\sigma^2 = \lambda E(w^2)
\]

Skewness and kurtosis are also simple functions of the expected values of powers of the weight \([11]\):
\[
\gamma_1 = \frac{E(w^3)}{\lambda^{1/2} E(w^2)^{3/2}}, \\
\gamma_2 = \frac{E(w^4)}{\lambda E(w^2)^2}.
\]  

In the special case that all weights are equal, skewness \( \gamma_1 = \lambda^{-1/2} \) and kurtosis \( \gamma_2 = \lambda^{-1} \) of the Poisson distribution are reproduced. For a narrow weight distribution, more precisely if \( E(w^2)^{1/2} \approx E(w^3)^{1/3} \approx E(w^4)^{1/4} \approx E(w) \), the two shape parameters are close to those of a Poisson distribution.

This observation suggests to approximate the CPD by a scaled Poisson distribution (SPD), \( f(x) \approx \tilde{f}(\tilde{x}) \) where \( \tilde{x} = s\tilde{n} \) and \( \tilde{n} \sim \mathcal{P}_\lambda \). The Poisson distributed number \( \tilde{n} \) is multiplied by a scaling factor \( s \) which depends on the weights such that the result \( \tilde{x} \) has a similar distribution as \( x \). To reproduce the first two moments of the CPD, \( E(\tilde{x}) = E(x) = \mu, \text{Var}(\tilde{x}) = \text{Var}(x) = \sigma^2 \), we select \( \tilde{\lambda} \) and the scaling factor \( s \) in the following way:
\[
\tilde{\lambda} = \frac{E(w)^2}{E(w^2)}, \\
s = \frac{E(w^2)}{E(w)}.
\]  

In applications, the expected values of the powers of the weight are not known. We have only a set of observed weights. We replace the expected values by the empirical mean values \( \bar{w}, \bar{w}^2 \), set \( \tilde{n} = n\bar{w}^2/\bar{w}^2 \) and use the observation that \( \tilde{n} \) is approximately Poisson distributed. The number \( \tilde{n} \) is called \textit{equivalent number of unweighted events} or \textit{effective number of events} because the \( n \) observed events provide the same statistical significance as \( \tilde{n} \) Poisson distributed events.
It can be shown that the SPD is a much better approximation of the CPD than a normal distribution \[41\]. This conclusion relies on simulations and on the analytic result that the SPD shape parameters are closer to the CPD values than those of the normal distribution where we have $\gamma_1 = \gamma_2 = 0$:

$$
\begin{align*}
\gamma_1(N) &= 0 < \gamma_1(SPD) \leq \gamma_1(CPD) , \\
\gamma_2(N) &= 0 < \gamma_2(SPD) \leq \gamma_2(CPD) .
\end{align*}
$$

(6.5) (6.6)

Two examples are shown in Fig. 6.1. In the left-hand graph the sum of $n$ exponentially distributed weights is displayed. The number $n$ is Poisson distributed with mean $20$. Even though the weight distribution is very wide, the SPD approximates the CPD (the discrete distribution has been approximated by a polygon) very well, much better than the normal approximation. In the right-hand graph the weight distribution is a truncated normal distribution $N(x > 0 | 1 , 1)$. Here the agreement is so good that the SPD and the CPD are hardly distinguishable.

Appendix 2: Error calculation of $(d − ct)$

**Poisson evaluation**

Given be two observed Poisson distributed numbers $N_1, N_2$, distributed according to $N_1 \sim \mathcal{P}_\lambda$ and $N_2 \sim \mathcal{P}_{a\lambda}$ with given $a$ and unknown $\lambda$. The log-likelihood of $\lambda$ is

$$
\ln L(\lambda) = -\lambda - a\lambda + N_1 \ln \lambda + N_2 \ln(a\lambda) .
$$

The value $\hat{\lambda}$ that maximizes $L$ obeys

$$
\frac{d \ln L}{d\lambda} = 0 = -(1 + a) + N_1/\hat{\lambda} + N_2/\hat{\lambda} .
$$

The result is

$$
\hat{\lambda} = \frac{N_1 + N_2}{1 + a} .
$$
With \( N_1 = d \), \( N_2 = \tilde{m} \), we get for the expected value \( \lambda = E(d) = E(ct) = E(cs\tilde{m}) \) and with \( E(\tilde{m}) = \lambda/(cs) \) we find

\[
\hat{\lambda} = \frac{d + \tilde{m}}{1 + 1/(cs)}.
\]

The error \( \delta \) of the difference \( d - ct = d - cs\tilde{m} \) is

\[
\delta^2 = \delta^2(d) + \delta^2(cs\tilde{m})
\]

\[
\hat{\lambda} = \delta^2(d) + c^2 s^2 \delta^2(\tilde{m}) .
\]

The errors have to be calculated from the expected values \( E(d) = \lambda, E(\tilde{m}) = \lambda/(cs) \):

\[
\delta^2 = \lambda (1 + cs)
\]

\[
\approx \frac{d + \tilde{m}}{1 + 1/(cs)} (1 + cs)
\]

\[
\approx cs(d + \tilde{m})
\]

\[
\approx c\sqrt{\frac{w^2}{w}} d + t .
\]

**Normal approximation**

Under the assumption that the expectation of \( d \) and \( ct \) is \( \lambda \), the errors are \( \delta^2_d = \lambda \) and \( \delta^2_t = \lambda/c \). We form a LS expression

\[
\chi^2 = \frac{(d - \lambda)^2}{\lambda} + \frac{(ct - \lambda)^2}{c\lambda}
\]

which we minimize with respect to \( \lambda \) :

\[
\frac{d\chi^2}{d\lambda} = -2\lambda \frac{d}{c\lambda^2} [c(d - \lambda) - (ct - \lambda)] - \frac{1}{c\lambda^2} [c(d - \lambda)^2 + (ct - \lambda)^2] = 0 .
\]

We find

\[
\hat{\lambda} = \left[ \frac{d^2 + ct^2}{1 + 1/c} \right]^{1/2}
\]

and

\[
\delta^2 = \delta^2_d + \delta^2_t = \hat{\lambda}(1 + c)
\]

Now we consider the more general case \( \delta^2_d = c'\lambda \).

\[
\chi^2 = \frac{(d - \lambda)^2}{c'\lambda} + \frac{(ct - \lambda)^2}{c\lambda}
\]

and get

\[
\hat{\lambda} = \left[ \frac{cd^2 + c'c^2d^2}{c + c'} \right]^{1/2}
\]

and

\[
\delta^2 = \delta^2_d + \delta^2_t = \hat{\lambda}(c' + c) .
\]
Appendix 3: Spline approximation

6.0.1 Quadratic \( b \)-splines

A quadratic \( b \)-spline (see Fig. 6.2) centered at \( x_0 \) with a bin width \( b \) is defined by

\[
B(x|x_0, b) = \begin{cases} 
\frac{1}{2b} \left( \frac{x - x_0 + 3/2b}{b} \right)^2 & \text{for } x_0 - 3b/2 \leq x \leq x_0 - b/2, \\
\frac{1}{2b} \left[ \frac{3}{2} - 2 \left( \frac{x - x_0}{b} \right)^2 \right] & \text{for } x_0 - b/2 \leq x \leq x_0 + b/2, \\
\frac{1}{2b} \left( \frac{x - x_0 - 3/2b}{b} \right)^2 & \text{for } x_0 + b/2 \leq x \leq x_0 + 3b/2, \\
0 & \text{else}.
\end{cases}
\]

For a given interval \([x_{\text{min}}, x_{\text{max}}]\) subdivided into \( n \) bins of width \( b = (x_{\text{max}} - x_{\text{min}})/n \) the number of quadratic \( b \)-spline functions is \( n+2 \). Two \( b \)-splines are centered outside the interval at \( x_0 = x_{\text{min}} - b/2 \) and \( x_{\text{max}} + b/2 \), respectively.

Cubic \( b \)-splines

Cubic \( b \)-splines are defined as follows:

\[
B(x|x_0, b) = \begin{cases} 
\frac{1}{6b} \left( 2 + \frac{x - x_0}{b} \right)^3 & \text{for } x_0 - 2b \leq x \leq x_0 - b, \\
\frac{1}{6b} \left[ -3 \left( \frac{x - x_0}{b} \right)^3 - 6 \left( \frac{x - x_0}{b} \right)^2 + 4 \right] & \text{for } x_0 - b \leq x \leq x_0, \\
\frac{1}{6b} \left[ 3 \left( \frac{x - x_0}{b} \right)^3 - 6 \left( \frac{x - x_0}{b} \right)^2 + 4 \right] & \text{for } x_0 \leq x \leq x_0 + b, \\
\frac{1}{6b} \left( 2 - \frac{x - x_0}{b} \right)^3 & \text{for } x_0 + b \leq x \leq x_0 + 2b, \\
0 & \text{else}.
\end{cases}
\]
For a given interval \( [x_{\text{min}}, x_{\text{max}}] \) subdivided into \( n \) bins of width \( b = (x_{\text{max}} - x_{\text{min}})/n \) the number of cubic \( b \)-spline functions is \( n + 3 \). Two \( b \)-splines are centered outside the interval at \( x_0 = x_{\text{min}} - b, \ x_{\text{max}} + b \) and two at the borders.

### Appendix 4: Choice of the bin width for parametric models

When we determine a parameter \( \theta \) of a distribution \( f(x|\theta) \) from a data sample of independent and identically distributed events \( (x_1, x_2, ..., x_N) \), the first choice of the method is a binning-free likelihood fit. Often this is not possible or too computer time consuming and we have to group the data in bins.

Wide bins have the advantage that the fluctuations of the number of entries can be approximated by a normal distribution, and that correlations due to the limited experimental resolution are small, but the disadvantage is that the parameter resolution can suffer. For instance the precision with which we can infer the location and width of a narrow peak over a smooth background decreases with increasing bin width. The exact relation between bin width and resolution depends on the shape of the distribution and the parameters of interest but we can derive a rule of thumb based on the Nyquist-Shannon theorem:

“If a function \( y(t) \) contains no frequency higher than \( \nu_{\text{max}} \), it is completely determined by giving its values at a series of points spaced \( 1/(2\nu_{\text{max}}) \) apart”

The Nyquist-Shannon theorem is of fundamental importance in the field of digital signal processing. It cannot be applied directly to our problem but it provides a hint for reasonable choices of the bin width. We can associate a maximum spatial frequency \( \nu_{\text{max}} \) to the function \( f(x) \) that we allow and apply the Nyquist-Shannon theorem to infer the number of points, here the number of histogram bins that we need to fix the function parameters. Instead of considering the frequency we turn to its inverse, the bandwidth \( w \) and get \( w < h_f/2 \) with \( w \) the bin width of the histogram and \( h_f \) the band width of the narrowest structure of the function \( f(x) \). Let us assume that the lowest bandwidth \( h_f \) is due to a narrow Gaussian peak with standard deviation \( \sigma \). The bandwidth of normal distribution can be approximated by \( h_f = \sqrt{2}\sigma \). (This is used in kernel density estimation.) With our crude estimate (a factor of two can hardly be argued), we find:

\[
w < \sigma/\sqrt{2}
\]

The bin width should not be larger than the standard deviation of the peak.

A quantitative result can be obtained in the following simple example:

A large data sample is used to determine the mean \( \mu \) and the standard deviation \( \sigma \) of a normal distribution. The data are histogrammed into bins of equal width \( b \). The nominal values of the distribution be \( \mu_0 = 0 \) and \( \sigma_0 = 1 \) and the data are contained in the interval \(-3 < x < 3 \). The LSF test statistic \( \chi^2 \) with \( B = 6/b \) bins is

\[
\chi^2 = \sum_{i=1}^{B} \frac{(cn_i - t_i)^2}{t_i}
\]

Here \( n_i \) is the number of events in bin \( i \) which, multiplied by a normalization constant \( c \), is compared to the integral of the normal distribution over the interval \( [x_{1i}, x_{2i}] \) given by the bin boundaries:

\(^1\)Unfortunately this definition is opposite to the definition of bandwidth in technical applications where \( \nu \) denotes the range of frequencies.
Fig. 6.3. Error estimates for the mean (dotted curve) and standard deviation (solid curve) of a normal distribution obtained from a data histogram. The errors are given in units of the error for infinitely small bins and the bin width is in units of the nominal standard deviation.

\[
\begin{align*}
t_i &= \int_{x_{i1}}^{x_{i2}} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right] dx \\
&= \text{erf}\left[\frac{x_{i2} - \mu}{\sqrt{2\sigma}}\right] - \text{erf}\left[\frac{x_{i1} - \mu}{\sqrt{2\sigma}}\right].
\end{align*}
\]

The minimum \(\chi^2 = 0\) of \(\chi^2\) is obtained with the settings \(c = 1\), \(\mu = 0\), \(\sigma = 1\), \(n_i = t_i\). The uncertainties of the parameters \(\mu\), \(\sigma\), correspond to the boundary \(\chi^2 = 1\). Of interest is the dependence of the errors \(\delta_\mu\) and \(\delta_\sigma\) on the bin width. To be independent of the number of events, we consider the ratios \(\delta_\mu/\delta_\mu_{\infty}\) and \(\delta_\sigma/\delta_\sigma_{\infty}\) where \(\infty\) stands for the limit of an infinite number of bins of zero width. In Fig. 6.3 the ratios are plotted as a function of the bin width in units of \(\sigma_0\). For a bin width equal to \(\sigma_0\) the resolution of \(\sigma\) is reduced by about 9.3\% with respect to the optimal resolution and for \(\mu\) the degradation is about 4.1\%. For the Nyquist estimate \(w = \sigma/\sqrt{2}\) the numbers are 4.8\% for \(\sigma\) and 2.1\% for \(\mu\) respectively.

Remark: This estimate does not apply to the non-parametric case of probability density estimation (PDE). There the quality of the agreement of the estimated distribution \(\hat{f}(x)\) which in our case is a histogram, is usually measured with the mean integrated square error MISE:

\[
\text{MISE} = \int_{-\infty}^{\infty} \left[\hat{f}(x) - f(x)\right]^2 dx.
\]

The formula for the MISE assumes a uniform distribution inside the bin. Here \(\hat{f}(x)\) corresponds to the lowest order spline approximation, while in physics applications the bin content is an estimate of the integral of the distribution and the quality of the agreement is mostly based on the goodness-of-fit statistics \(\chi^2\).

Appendix 5: Simplex convergence

Simplex is a wide spread robust method to find the parameters that maximizes of a non-linear function. Contrary to many other minimization methods it does not require to determine derivatives.
of the function that has to be maximized. It has been shown that Simplex converges in deterministic problems but not necessarily in problems where the data are affected by noise. The problem increases with the number of parameters.

A description of the simplex algorithm can be found in many textbooks and need not be repeated here. There are many variants (see refs. in [43] of the original version of Nelder and Mead ([42]). Standard Simplex [42] has been used in all fits of this manuscript. With \( N \) parameters, we initialize the first of the \( N + 1 \) parameters vectors (points in the \( N \)-dimensional parameter space) to the true parameter values known in our Monte Carlo studies. The remaining vectors differ each in one parameter value by the expected uncertainty from the true value. If the number of parameters is large, and especially if the parameters are correlated, Simplex fits have the tendency to stop without having reached the function minimum [43]. Simplex may choose shrinkage while a reflection of the worst parameter point could be the optimal choice. Finally, all points have almost equal parameter coordinates such that the convergence criterion is fulfilled. Further improvement steps are so small that reducing the convergence parameter does not change the result. The convergence problem is studied in great detail in [43] and a solution which introduces stochastic elements in the stepping process is proposed.

In this manuscript a different approach is followed. After Simplex signals convergence, the fit is repeated where the best point so far obtained is kept and the remaining points are initialized in the same way are before. Alternatively these points are chosen randomly centered at the best value.

Both ways have been applied. The following parameters were used: The maximum number of steps in Simplex was set to a rather low value of 5000. The starting values of the parameters were set to the true values which are known for the Monte Carlo studies. The convergence parameter was \( 10^{-12} \). If the log-likelihood values of two subsequent steps differ by less than this value, the fit is stopped. The fit is repeated up to 200 times each time keeping the best parameter point but modifying for each of the remaining \( N \) points one parameter. Then a maximum of 1000 or sometimes 10000 additional fits with random initialization of the \( N \) points are added. If 20 subsequent fits do not change the double precision function value within the precision of the computer, the fitting is terminated. The fit with curvature penalty is especially problematic. This is due to the strong correlation of the neighboring bin contents. With entropy and norm penalties usually 50 repetitions are sufficient.

Appendix 6: Expectation Maximization Algorithm

The EM method finds iteratively the MLE in situations where the statistical model depends on latent variables. The method goes back to the sixties, has been invented several times and has been made popular by Dempster, Laird and Rubin [34]. A very comprehensive introduction to the expectation maximization (EM) algorithm is given in the Wikipedia article Ref. [44]. The EM algorithm exists in many different variants. We will restrict our discussion to its application to classification problems.

To get an idea of the method, we consider a simple standard example. Let us assume that we have a sample of observations \( x_1, \ldots, x_N \) each drawn from one of \( M \) overlapping normal distribution \( f_m(x|\mu_m) \sim \mathcal{N}(\mu_m, s) \) with unknown mean values \( \mu_1, \ldots, \mu_M \) and given standard deviation \( s \). The log-likelihood for the parameters is

\[
\ln L(\mu_1, \ldots, \mu_N) = \sum_{m=1}^{M} \frac{\sum_{i=1}^{N} z_{mi} x_i}{\sum_{i=1}^{N} z_{mi}} - \mu_m
\]

where the classification variable \( z_{mi} = 1 \) if \( x_i \) belongs to the normal distribution \( m \) and \( z_{mi} = 0 \) otherwise. If we know the classification variables, we get the MLE of the parameter \( \mu_m \):

\[
\hat{\mu}_m = \frac{\sum_{i=1}^{N} z_{mi} x_i}{\sum_{i=1}^{N} z_{mi}}.
\]
If this is not the case, we can estimate $z_{mi}$ from the observed distribution. In the EM formalism $z_{mi}$ is called missing or latent variable. We can solve our problem iteratively with two alternating steps, an expectation and a maximization step. We start with a first guess $\mu_m^{(1)}$ of the parameters of interest and estimate the missing data. In the expectation step $k$ we compute the probability $g_{mi}^{(k)}$ that $x_i$ belongs to subdistribution $m$. It is proportional to the value of the distribution $f_m(x_i|\mu_m)$ at $x_i$:

$$g_{mi}^{(k)} = \frac{f_m(x_i|\tilde{\mu}_m^{(k)})}{\sum_{j=1}^{M} f_j(x_i|\tilde{\mu}_m^{(k)})}.$$ 

The probability $g_{mi}$ is the expected value of the latent variable $z_{mi}$. The expected log-likelihood is

$$Q(\mu, \tilde{\mu}^{(k)}) = \sum_{m=1}^{M} \left( \sum_{i=1}^{N} g_{mi}^{(k)} x_i - \mu_m \right).$$

In the maximization step we obtain the MLEs

$$\tilde{\mu}_m^{(k+1)} = \sum_{i=1}^{N} g_{mi}^{(k)} x_i,$$

which are used in the following expectation step. The iteration converges to the overall MLE.

Let us generalize this procedure. Given be a probability distribution $p(x, z|\theta) = g(z|x, \theta)p_1(x|\theta)$ depending on a parameter vector $\theta$ and a sample of observations $x_1, ..., x_N$. The distribution $g$ of the latent variables $z$ is a function of $\theta$ and $x$.

- **Expectation step**: For the parameter vector $\theta^{(k)}$ we compute the distribution $g$ of the hidden variables $z$. We form the log-likelihood function $\ln L(\theta|x, z)$ which is a random variable as it depends on the random $z$. Averaging over $z$, we obtain the expected value $Q(\theta, \tilde{\theta}^{(k)})$ of the log-likelihood:

$$Q(\theta, \tilde{\theta}^{(k)}) = E_{z|x, \theta^{(k)}} [\ln L(\theta|x, z)] .$$

The conditional expectation means that we average over $z$ given the distribution of $g(z)$ obtained for fixed values $x$ and $\tilde{\theta}^{(k)}$:

$$Q(\theta, \tilde{\theta}^{(k)}) = \int_z \ln L(\theta|x, z) g \left( z|x, \tilde{\theta}^{(k)} \right) dz .$$

If the values of the vector components $z$ are discrete, the integral is replaced by a sum over all $J$ possible values:

$$Q(\theta, \tilde{\theta}^{(k)}) = \sum_{j=1}^{J} \ln L(\theta|x, z_j) g \left( z_j|x, \tilde{\theta}^{(k)} \right) .$$

Alternatively, somewhat less efficient, we can insert the expected values of the latent variables:

$$Q(\theta, \tilde{\theta}^{(k)}) = \ln L(\theta|x, E(z)) .$$

- **Maximization step**: The MLE $\theta^{(k+1)}$ is computed:

$$\theta^{(k+1)} = \arg \max_\theta Q(\theta|\theta^{(k)}) .$$

The procedure is started with a first $\theta^{(1)}$ guess of the parameters and iterated. It converges to a minimum of the log-likelihood. To avoid that the iteration is caught by a local minimum, different starting values can be selected. It is especially useful in classification problems in connection with p.d.f.s of the exponential family, where the maximization step is relatively simple.

To the exponential family belong among others the normal, Poisson, exponential, gamma, chi-squared distribution.
Example 26. Unfolding a histogram

Experimental data are collected in form of a histogram with $N$ bins. The number of events in bin $i$ be $d_i$. The experiment suffers from an imperfect resolution and from acceptance losses which we have to correct for. The "true" histogram with $M$ bins contains $\theta_j$ events in bin $j$. Knowing the measurement device we can simulate the experimental effects and determine the matrix $A$ which relates $\theta$ with the expected values of the numbers $d$: $E(d_i) = \sum_{j=1}^{M} A_{ij} \theta_j$. The element $A_{ij}$ is the probability to observe an event in bin $i$ which belongs to the bin $j$ in the true histogram. The missing information is the number of events $d_{ij}$ in an observed bin $i$ that belong to the true bin $j$. Hence there are $M$ missing variables per bin. The number $d_{ij}$ is Poisson distribution with mean $A_{ij} \theta_j$. The likelihood depends only on the hidden variables:

$$\ln L(\theta|d_{11}, ..., d_{NM}) = \sum_{j=1}^{M} \sum_{i=1}^{N} [-A_{ij} \theta_j + d_{ij} \ln A_{ij} \theta_j].$$

The following alternating steps are repeated:

- **Expectation step**: We have

$$Q(\theta, \hat{\theta}_j^{(k)}) = E_{d_{ik}} \ln L = \sum_{j=1}^{M} \sum_{i=1}^{N} [-A_{ij} \theta_j + E(d_{ij}^{(k)}) \ln A_{ij} \theta_j].$$

The expected value $E(d_{ij}^{(k)})$ conditioned on $d_i$ and $\hat{\theta}_j^{(k)}$ is given by $d_i$ times the probability that an event of bin $i$ belongs to true bin $j$.

$$E(d_{ij}^{(k)}) = d_i \frac{A_{ij} \hat{\theta}_j^{(k)}}{\sum_{j=1}^{M} A_{ij} \hat{\theta}_j^{(k)}}.$$

We get

$$Q(\theta, \hat{\theta}_j^{(k)}) = \sum_{j=1}^{M} \sum_{i=1}^{N} [-A_{ij} \theta_j + \frac{A_{ij} \hat{\theta}_j^{(k)}}{\sum_{m=1}^{M} A_{im} \hat{\theta}_m^{(k)}} \ln A_{ij} \theta_j].$$

- **Maximization step**:

The computation of the maximum of $Q$ is easy, because the components of the parameter vector $\theta$ appear in independent summands.
\[
\frac{\partial Q}{\partial \theta_j} = \sum_{i=1}^{N} \left[ -A_{ij} + d_i \frac{A_{ij} \hat{\theta}_j^{(k)}}{\sum_{j=1}^{M} A_{ij} \hat{\theta}_j^{(k)}} \right] = 0 ,
\]

\[
\hat{\theta}_j^{(k+1)} \sum_{i=1}^{N} A_{ij} = \sum_{i=1}^{N} d_i \frac{A_{ij} \hat{\theta}_j^{(k)}}{\sum_{j=1}^{M} A_{ij} \hat{\theta}_j^{(k)}} ,
\]

\[
\theta_j^{(k+1)} = \sum_{i=1}^{N} d_i \frac{A_{ij} \theta_j^{(k)}}{\sum_{j=1}^{M} A_{ij} \theta_j^{(k)}} / \alpha_j .
\]

\[\Sigma_{i=1}^{N} A_{ij} = \alpha_j\] is the average acceptance of the events in the true bin \(j\).

References

1. W. H. Richardson, *Bayesian-Based Iterative Method of Image Restoration*, J. Opt. Soc. Am 62 (1972) 55.
2. L. B. Lucy, *An iterative technique for the rectification of observed distributions*, Astron. J. 79 (1974) 745.
3. L. A Shepp, Y. Vardi, *Maximum Likelihood Reconstruction for Emission Tomography*, IEEE transactions on Medical Imaging 1 (1982) 113.
4. Y. Vardi, L. A. Shepp and L. Kaufmann, *A statistical model for positron emission tomography*, J. Am. Stat. Assoc. 80 (1985) 8.
5. H. N. Mühlthi and B. Schorr, *On an iterative method for the unfolding of spectra*, Nucl. Instr. and Meth. A257 (1987) 371, H. N. Mühlthi, B. Schorr, W. Törnig *On properties of the iterative maximum likelihood reconstruction method*, Math. Meth. Appl. Sci. 11 (1989) 331.
6. R. Narayan, R. Nityananda, *Maximum Entropy Image Restoration in Astronomy*, Ann. Rev. Astron. and Astrophys. 24 (1986) 127.
7. A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Sov. Math. 5 (1963) 1035.
8. V. Blobel, *Unfolding methods in high-energy physics experiments*, CERN Yellow Report 85-09 (1985) 88.
9. V. B. Anykeev, A. A. Spiridonov and V. P. Zhigunov, *Comparative investigation of unfolding methods*, Nucl. Instr. and Meth. A303 (1991) 350.
10. M. Schmelling, *The method of reduced cross-entropy - a general approach to unfold probability distributions*, Nucl. Instr. and Meth. A340 (1994) 400.
11. L. Lindemann and G. Zech, *Unfolding by weighting Monte Carlo events*, Nucl. Instr. and Meth. A354 (1995) 516.
12. G. D’Agostini, *A multidimensional unfolding method based on Bayes’ theorem*, Nucl. Instr. and Meth. A 362 (1995) 487.
13. G. Zech, *Comparing statistical data to Monte Carlo simulation - parameter fitting and unfolding*, DESY Report 95-113 (1995).
14. A. Hoecker and V. Kartvelishvili, *SVD approach to data unfolding*, Nucl. Instr. and Meth. A 372 (1996), 469.
15. M. C. Abreu et al. A 4-dimensional deconvolution method to correct Na38 experimental data, Nucl. Instr. and Meth. A 405 (1998) 139.
16. P. Magan, F. Courbin and S. Sohy, Deconvolution with correct sampling, Astrophys. J. 494 (1998) 472.
17. G. D’Agostini, Improved iterative Bayesian unfolding, arXiv:1010.0632v1 (2010).
18. N. Milke et al. Solving inverse problems with the unfolding program TRUE: Examples in astroparticle physics, Nucl. Instr. and Meth. A 697 (2013) 133.
19. G. Zech, Iterative unfolding with the Richardson-Lucy algorithm, Nucl. Instr. and Meth. A 716 (2013) 1.
20. H. P. Dembinski, M. Roth, An algorithm for automatic unfolding of one-dimensional distributions, Nucl. Instr. and Meth. A 729 (2013) 725.
21. V. Blobel, Unfolding, In Data Analysis in High Energy Physics, ed. O. Behnke et al., Wiley-VCH, Weinheim (2013) 187.
22. M. Kuusela and V. M. Panaretos, Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification, Annals of Applied Statistics 9 (2015) 1671.
23. M. Kuusela, Statistical Issues in Unfolding Methods for High Energy Physics, Master’s thesis, Aalto University, Finland (2012).
24. I. Volobouev, On the Expectation-Maximization Unfolding with smoothing, arXiv:1408.6500v2 (2015).
25. G. Zech, Introduction to Statistics and Data Analysis for Physicists, Verlag Deutsches Elektronens-Synchrotron, http://www-library.desy.de/elbook.html (2014).
26. I. Volobouev, On the Expectation-Maximization Unfolding with smoothing, arXiv:1408.6500v2 (2015).
27. G. Cowan, A survey of unfolding methods for particle physics, G. Cowan, Statistical Data Analysis, Oxford University Press (1998).
28. D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley, New York (1992).
29. I. Narsky, F. C. Porter, Statistical Analysis Techniques in Particle Physics, Wiley-VCH, Berlin (2013) 89.
30. G. Bohm, G. Zech, Introduction to Statistics and Data Analysis for Physicists, Verlag Deutsches Elektronen-Synchrotron, http://www-library.desy.de/elbook.html (2014).
31. J. A. Nelder and R. Mead, A simplex method for function minimization, The Computer Journal, 7 (1965) 308.
43. J. J. Tomik, *On Convergence of the Nelder-Mead Simplex algorithm for unconstrained stochastic optimization*, PhD Thesis, Pennsylvania State University, Department of Statistics (1995).
44. https://en.wikipedia.org/wiki/Expectation-maximization-algorithm.
