Unfolding methods in high-energy physics experiments

V. BLOBEL

II.Institut für Experimentalphysik der Universität Hamburg

ABSTRACT. Distributions measured in high-energy physics experiments are often distorted or transformed by limited acceptance and finite resolution of the detectors. The unfolding of measured distributions is an important, but due to inherent instabilities a very difficult problem. Methods for unfolding, applicable for the analysis of high-energy physics experiments, and their properties are discussed. An introduction is given to the method of regularization.

Nay, answer me: stand, and unfold yourself
- William Shakespeare, *from Hamlet*

You can get it wrong and still you think it's all right
- John Lennon and Paul McCartney, *from*
We can work it out

1. INTRODUCTION

One of the objectives of experimental physics is to measure distributions $f(x)$ of some physical variable x. Examples from high-energy physics are angular distributions, invariant-mass spectra and structure functions. The set $\{x\}$ of values measured in particle reactions, or events, can be regarded as a random sample, drawn from a distribution $f(x)$ of a one-dimensional random variable x, and it is the purpose of data analysis, to make inferences about $f(x)$ from the set $\{x\}$. The measured distribution $\hat{f}(x)$[1], usually determined in the form of a histogram [1], differs from the true distribution $f(x)$ by statistical errors $\epsilon(x)$; it can be used to test theoretical predictions $f_{th}(x)$. In the case of an expression $f_{th}(x,a)$, depending on parameters a, values for the parameters can be extracted by fitting $f_{th}(x,a)$ to the data $\hat{f}(x)$.

In high-energy physics experiments often the quantity x and the distribution of x cannot be measured directly, due to the imperfection of the detector. Two detector effects can be distinguished, limited acceptance and resolution. Limited acceptance means that the probability to observe a given event is less than one; the acceptance may depend on the kinematical region. The second effect, limited resolution, means that the quantity x in a given event cannot be determined exactly, but it can only be measured with a certain measurement error. Both effects result in a distortion of the measured distribution which can be expressed in the following way: instead of the physically relevant variable x, a variable y and its distribution $g(y)$ is measured. Due to a transformation property of the measurement process, the measured variable y may differ from x completely; for a one-dimensional variable x the variable y may be

[1]Measured quantities, and quantities, derived from measured quantities, are denoted by a hat, for example $\hat{f}(x)$.

multidimensional. The transformation may also include additional kinematical effects, for example in scattering experiments there may be effects from radiation and Fermi motion (in case of heavy targets) [2].

In this paper only the case is considered, where the distributions $f(x)$, defined over the range $a \leq x \leq b$, and $g(y)$ are related by the convolution integral

$$g(y) = \int_a^b A(y,x) f(x)\, dx\,. \tag{1.01}$$

In the theory of integral equations equation (1.01) is called the Fredholm integral equation of first kind. Integral equations occur in many branches of computational physics. Examples from different fields of science are given in [3] and in [4], which also contains many references. The function $A(y,x)$, called a kernel in the theory of integral equations, describes the response of the detector including the transformation from x to y. For a given $x = x_0$, the response of the detector in the variable y is $A(y,x_0)$. In practice the distribution $\hat{g}(y)$ actually measured differs from the expected distribution $g(y)$ by statistical errors $\epsilon(y)$:

$$\hat{g}(y) = \int_a^b A(y,x) f(x)\, dx + \epsilon(y) \tag{1.02}$$

An accurate determination of the response function $A(y,x)$ is essential for any meaningful analysis of the data. Information on the behaviour of the detector can be obtained by test measurements with a known $f(x)$. For example a hadron calorimeter can be exposed to a particle beam with well known fixed energy $x = x_0$, then $f(x) = \delta(x - x_0)$ and the result of the measurement of the energy y in the calorimeter is directly $A(y,x_0)$:

$$\int_a^b A(y,x)\delta(x - x_0)\, dx = A(y,x_0). \tag{1.03}$$

Complex high-energy experiments usually require Monte-Carlo calculations of the detector response. This method allows the determination, for a given dependence $f(x)$, of the expected distribution $g(y)$ of any measurable quantity y by a detailed simulation of the measuring process in the detector. Often detector data are generated in the same format as real data, and can be processed by an identical chain of reconstruction and analysis programs.

If the effect of limited resolution is negligible and the measured distributions are essentially affected only by limited acceptance, the correction for these effects is usually not very difficult. There are two possibilities for the data analysis in case of a large distorting effect due to limited resolution. If a specific theoretical prediction $f_{th}(x)$ is to be compared with the data, the corresponding expected distribution $g_{th}(y)$ can be evaluated by equation (1.01). If the distribution $g_{th}(y)$ agrees with the measured distribution $\hat{g}(y)$ within statistical errros, one can conclude, that the tested prediction gives a consistent description of the measured process. In the case of disagreement however a publication of a detector-dependent measured distribution $\hat{g}(y)$ is of little use. The comparison with alternative theoretical predictions is impossible, unless they are also convoluted with the detector response using equation (1.01). The other possibility is the reconstruction of $f(x)$ from the measured distribution $\hat{g}(y)$.

The reconstruction of $f(x)$ from the measured distribution $\hat{g}(y)$ is called unfolding, and it is a statistical estimation problem. Unfolding is a complicated problem, because, in mathematical classification, it is an ill-posed problem [3], which can have wildly oscillating solutions [4]. In spite of its practical importance, it is not mentioned in standard textbooks on statistics, with only few exceptions [5]; in general many misconceptions exist and are used in practical applications. It is characteristic of these methods, that they are usually described only in words; the standard criteria for statistical estimation methods, the requirements for efficiency, consistency and unbiasedness are usually difficult to discuss for these methods.

One example of the combined effects of acceptance, transformation and resolution in a high-energy physics experiment is the measurement of the cross section $d\sigma/d\eta$ as a function of the inelasticity $\eta = (E_h - m_p)/E_\nu$ in neutral current reactions

$$\nu N \rightarrow \nu X$$

in a narrow band neutrino beam[1] [6]. The detector allows the measurement of the energy E_h of the hadronic system X above a threshold of a few GeV with a certain resolution, and in addition the distance r of the interaction from the beam axis can be measured (m_p is the proton mass). The beamflux $\Phi(E_\nu, r)$ has a two-peak structure at a fixed distance r, since the neutrinos have their origin in decays of π- and K-Mesons. Clearly the measured data do not allow the reconstruction of the value η of the inelasticity in individual events, since the neutrino energy E_ν is not known. The knowledge of the beamflux $\Phi(E_\nu, r)$ together with the known acceptance and resolution in E_h and r allows the calculation for any given $d\sigma/d\eta$ of the resulting distribution of the events in the (E_h, r)-plane. The transformation between η and (E_h, r) is completely defined and can be expressed by the formula

$$g(E_h, r) = \int A(E_h, r, \eta) \frac{d\sigma}{d\eta} d\eta, \tag{1.04}$$

which has the same structure as equation (1.01). Unfolding in this case means the reconstruction of $d\sigma/d\eta$ from the measured distribution $g(E_h, r)$.

Another example from high energy physics is the measurement of the total cross section for $\gamma\gamma$-interactions, which is possible with e^+e^- storage rings at high energies by a measurement of the reaction [7] [8]

$$e^+e^- \rightarrow e^+e^-X.$$

Because some fraction of particles of the hadronic final state X are emitted at small angles with respect to the beams, they escape detection by the detector and the measured invariant mass of the hadronic system is on average smaller than the true invariant mass. A model of the physical process is necessary in this case for the calculation of the response function $A(y, x)$. Model parameters have to be determined from the data.

[1] The usual symbol y for the inelasticity is replaced here by the symbol η, to avoid confusion with measured variables, denoted by the symbol y in this paper.

If the variables x and y are discrete variables, the integral in equation (1.01) has to be replaced by a sum. A discrete set can always be mapped on a set of consecutive integers, and therefore one can assume, that the random variables x and y have ranges $1 \ldots m$ and $1 \ldots n$, respectively. Instead of functions $f(x)$ and $g(y)$ of continuous variables, there is a finite number of elements f_j, $j = 1 \ldots m$ and g_i, $i = 1 \ldots n$, and the convolution equation can be written in the form

$$\hat{g}_i = \sum_{j=1}^{m} A_{ij} f_j + \epsilon_i \qquad \text{or short} \qquad \hat{g} = Af + \epsilon, \tag{1.05}$$

where f is a n-vector, $\hat{g}$ and ϵ are m-vectors, and A is a m-n matrix. The equation can be interpreted as a discrete approximation of the integral equation (1.02). Any numerical solution of the integral equation for continuous variables will require an approximation by a finite number of elements. The simplest discretization is the representation of the distributions by histograms f and $\hat{g}$ (a more general discretization method is discussed in chapter 4.2).

As mentioned before, the correction is not very difficult, if the dominant effect of the detector is limited acceptance. In this case all elements A_{ij} are zero for $i \neq j$ (with $n = m$) and only the elements A_{jj}, the acceptance probability for bin j, have to be considered. A common method for pure acceptance correction is the following: Monte Carlo events with x-values generated according to some fixed assumption $\bar{f}$ are processed, simulating the detector response, and the accepted events are used to fill a histogram $\bar{g}$. The binwise ratio $\bar{g}_j/\bar{f}_j$ of the histograms gives the values A_{jj} of the acceptance probabilities for bins j. The correction of the measured bin contents $\hat{g}_j$ is then made according to

$$\hat{f}_j = \hat{g}_j \left(\frac{\bar{f}_j}{\bar{g}_j} \right), \tag{1.06}$$

to obtain corrected bin contents $\hat{f}_j$. Since $A_{ij} = 0$ for $i \neq j$, the correction factor does not depend on the used MC-input histogram $\bar{f}$. There are no difficulties or problems with this method, which may be called the factor method, except perhaps if the acceptance probability changes rapidly within a few bins [9].

The situation is completely different, if a correction for limited resolution becomes necessary. The principal difficulty of unfolding is easily demonstrated. For the discrete case with $n = m$ (square matrix A), the straightforward application of standard analysis methods suggests the solution

$$\hat{f} = A^{-1} \hat{g}, \tag{1.07}$$

where A^{-1} is the matrix inverse to A; this method may be called inversion method. The expectation value $E(\hat{f})$ of the estimate $\hat{f}$ is equal to the true f, provided the measured $\hat{g}$ has no bias, i.e. $E(\hat{g}) = Af$:

$$E(\hat{f}) = E(A^{-1}\hat{g}) = A^{-1} E(\hat{g}) = A^{-1} A f = f. \tag{1.08}$$

An estimate with this desirable property is called consistent. Error propagation yields

$$V(\hat{f}) = A^{-1} V(\hat{g}) A^{-1} \tag{1.09}$$

for the covariance matrix $V(\hat{f})$, calculated from the covariance matrix $V(\hat{g})$ of the measured data. This method is probably tried first by many people, when confronted with an unfolding problem. The result however is often very disappointing, as shown in the numerical example below, and it is understandable that people after trying this method turn to a heuristic method which provides 'better' results.

Example: Unfolding of a distribution of a discrete variable. The case $n = m = 20$ is assumed, with the following response matrix:

$$A = \begin{pmatrix} 0.75 & 0.25 & 0 & & \\ 0.25 & 0.50 & 0.25 & 0 & \\ 0 & 0.25 & 0.50 & 0.25 & 0 \\ & 0 & 0.25 & 0.50 & 0.25 \\ & & 0 & 0.25 & 0.50 \\ & & & & \ddots \end{pmatrix}$$

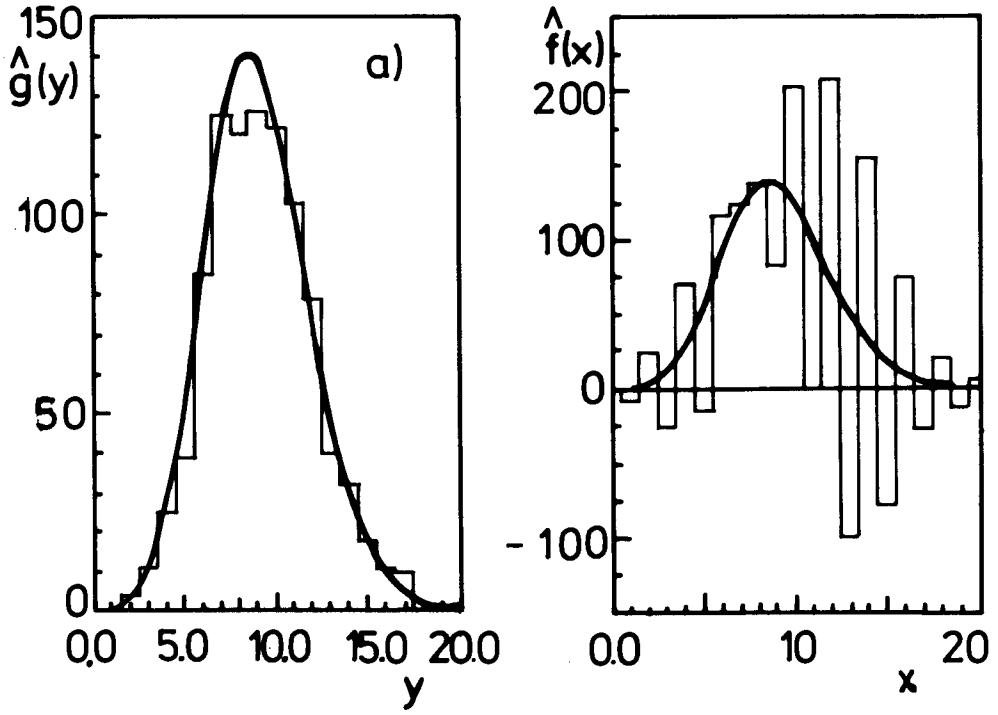

Figure 1. Distribution of the measured quantity y (a) and oscillating result of unfolding (b) using equation (1.05), shown as histograms. The original dependence is shown as a curve in both cases.

The probability to observe for y the true value x is 50 % (except for $x = 1$ and $x = n$). A certain dependence $f(x)$ is assumed and the expected distribution $g(y)$ is calculated by $g = A f$; a random sample of 1000 independent measurements of the variable y is generated. The result of the simulation is shown in Figure 1a as a histogram. Equation (1.07) is used to obtain an estimate $\hat{f}$ for the original distribution, and this is

shown as a histogram in Figure 1b together with the original distribution (smooth curve). The result has an oscillating behaviour, and the covariance matrix shows large negative correlations, especially between adjacent values, although the effect of the resolution on the shape of the measured distribution appears to be rather small. The result of this example is typical for the direct solution of the unfolding problem, based on equation (1.07). Clearly this method is not acceptable.

In several experiments the factor method, as discussed above for the case of a pure acceptance correction, is used in the case of limited resolution as well. However now because of $A_{ij} \neq 0$ for $i \neq j$ the correction factor according to formula (1.06) will depend on the MC input histogram $\bar{f}$, and will only be correct, if $\bar{f}$ is the true histogram. Since this is not known (otherwise the measurement would not be necessary), some assumption is necessary and the corrected values $\hat{f}_j$ will be biased towards the MC input histogram. Often in applications of this kind, the MC input histogram is adjusted in several iterations, to get a histogram $\bar{g}$, which describes the measured histogram $\hat{g}$ well. The essential point in applications of this method is to use always a smoothed histogram $\bar{f}$, otherwise the repeated application of equation (1.06) can be shown to give (in case of convergence) the result of the inversion method (which exactly reproduces $\hat{g}$). In general one can say, that the factor method applied to the case of finite resolution gives a biased result, with the danger to underestimate statistical and systematic uncertainties. Quite often the corrected result is presented in the form of a histogram with many bins, which are narrower than the resolution, and which cannot be resolved by the detector.

Acceptable unfolding results can be obtained by regularization methods [10] [11], which suppress the spurious oscillatory component in the solution. Regularization can be interpreted as the use of certain a-priori information on the degree of smoothness of the true solution. Since this can introduce a bias, the weight of the a-priori information has to be determined by statistical methods in order to keep any possible bias small compared to statistical errors. One unavoidable consequence of acceptable unfolding is the limitation of the number of unfolded data points, according to the resolution and the statistical accuracy. A measurement with limited resolution always means a loss of statistical accuracy.

A special feature of high-energy physics experiments is the fact that the response function is usually known only implicitly by the simulation of particle reactions in the detector. Limited statistical accuracy of the often very time-consuming Monte-Carlo calculations can introduce systematic uncertainties. The usual requirement 'number of MC events $\approx$ number of measured events' is insufficient. Even more difficult, the Monte-Carlo simulation may require some assumptions, which have to be tested by comparison with the data, and therefore unfolding methods should allow sensitive tests of the assumptions. In addition the discretization has to be done rather carefully to avoid a further deterioration of the already limited resolution.

In this paper a general unfolding method, based on fundamental statistical principles is discussed in detail together with numerical examples. The method is similar to methods, used in other fields of science [12] [13] [14] and allows the inclusion of certain a priori information (regularization). It has been applied to several high-energy physics

experiments in an earlier [6] [15] and in the present version [7] [8]. The discussion is restricted to the case of a one-dimensional variable x and its distribution; the measured variable however may be multidimensional.

The description and discussion of the unfolding method is preceeded by two chapters which introduce the basic concepts of statistics including parameter estimation and the parametrization of functions by spline and orthogonal functions.

2. STATISTICS

2.1 ONE-DIMENSIONAL RANDOM VARIABLES

The result of a measurement can be characterized by one or more real numbers x_i, $i = 1 \ldots$. The probability that an experiment yields a result $a \leq x < b$ is given by

$$P(a \leq x < b) = \int_a^b f(x)\,dx \tag{2.01}$$

for a continuous random variable (r.v.) x, where $f(x)$ is the probability density function of the variable x. A probability density function is a nonnegative function with unit integral:

$$f(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{2.02}$$

Physicists usually call a probability density function (p.d.f.) a distribution, in statistics this name is reserved for the integrated p.d.f., which may be called cumulative distribution $F(x)$:

$$F(x) = \int_{-\infty}^{x} f(x')\,dx' \tag{2.03}$$

with $F(-\infty) = 0$ and $F(\infty) = 1$. An important parameter characterising the location of a random variable x is the expectation value of x, denoted by $E(x)$ and defined by

$$E(x) = \int_{-\infty}^{\infty} x\,f(x)\,dx. \tag{2.04}$$

Generalized for an arbitrary function $h(x)$, the expectation value of $h(x)$ is defined by

$$E(h) = \int_{-\infty}^{\infty} h(x) f(x)\,dx. \tag{2.05}$$

The expectation values of x^n and of $(x - E(x))^n$ are called n-th algebraic moments μ_n and n-th central moments μ_n', respectively. The expectation value of x, or mean value μ of x, is equal to the first algebraic moment μ_1,

$$\mu = \mu_1 = E(x). \tag{2.06}$$

The second central moment μ_2' is a measure of the spread of the distribution, it is called variance $V(x)$ and its definition is

$$V(x) = E(\,(x - E(x))^2\,). \tag{2.07}$$

The variance is abbreviated by σ^2, and σ is called the standard deviation.

The normal distribution. The normal or gaussian distribution is in practice the most important distribution, since measurement errors often follow the normal distribution. The density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad x \in (-\infty, \infty). \tag{2.08}$$

The normal distribution has two parameters μ and σ with $E(x) = \mu$ and $V(x) = \sigma^2$. The probability for x to fall into the region $\mu \pm \sigma$ is 68.3 %. The normal distribution for the case $\mu = 9$ and $\sigma = 3$ is shown in Figure 2 as a curve.

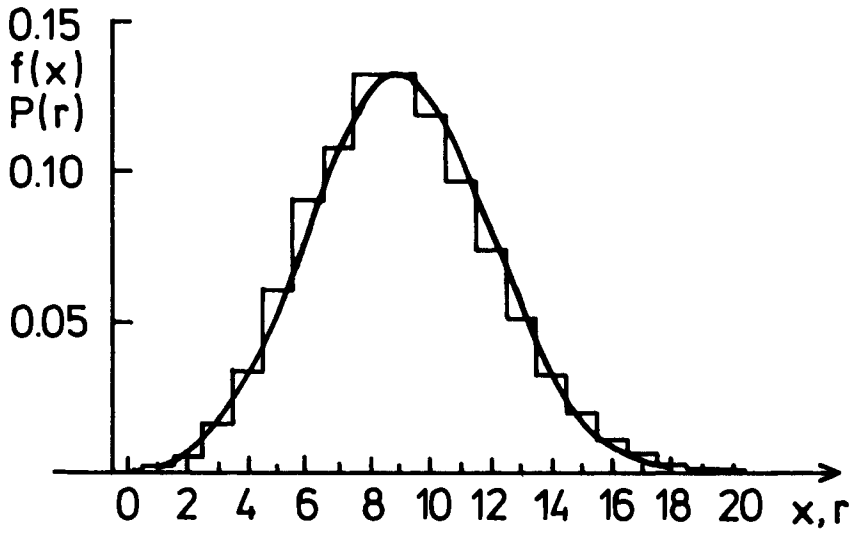

Figure 2. Normal distribution with $\mu = 9$ and $\sigma = 3$, shown as curve, and Poisson distribution with $\mu = 9$, shown as histogram.

A linear function $y = ax + b$ of a random variable x is again a random variable. The expectation value and the variance are

$$\mu_y = a\mu_x + b \quad \text{and} \quad \sigma_y^2 = a^2\sigma_x^2. \tag{2.09}$$

For a general transformation $y = h(x)$ the p.d.f. of y is

$$g(y) = \int_{-\infty}^{\infty} \delta(y - h(x)) f(x)\, dx = \frac{f(x)}{|h'(x)|}\ , \tag{2.10}$$

if the function $y = h(x)$ is a one-to-one function.

In the case of a discrete variable r possible values can always be represented by a set of consecutive integers, $r \in \{a, a+1, a+2 \ldots, b-1, b\}$. The probability, that an experiment yields the result r is denoted by the nonnegative number $P(r)$. Expectation values are defined in analogy to the case of continuous random variables, replacing integrals by sums.

The Poisson distribution. If events occur at a constant rate, the probability of observing in a given time interval exactly r events, is given by the Poisson distribution. For a Poisson distribution,

$$P(r|\mu) = e^{-\mu}\frac{\mu^r}{r!} \qquad r \in \{0, 1 \ldots \infty\} \tag{2.11}$$

represents the probability of observing r events, if the mean value is μ. The Poisson distribution has only one parameter μ, with $E(r) = \mu$ and $V(r) = \mu$. The Poisson distribution with $\mu = 9$ is shown in Figure 2 as a histogram. A comparison shows, that

for not too small mean values (say $\mu \geq 5$) the normal distribution represents a good approximation to the Poisson distribution, except in the tails.

2.2 MULTIDIMENSIONAL RANDOM VARIABLES

For a twodimensional random variable with components x and y the probability to observe (x, y) with $a \leq x < b$ and $c \leq y < d$ is

$$P(a \leq x < b, c \leq y < d) = \int_c^d \int_a^b f(x,y)\, dx\, dy \tag{2.12}$$

with the twodimensional p.d.f. $f(x, y)$, which obeys the normalization condition

$$\iint f(x,y)\, dxdy = 1 \tag{2.13}$$

with integration limits $-\infty$ and $+\infty$. Projections of the distribution $f(x, y)$ are called marginal distributions:

$$f_x(y) = \int_{-\infty}^{\infty} f(x,y)\, dx \qquad f_y(x) = \int_{-\infty}^{\infty} f(x,y)\, dy. \tag{2.14}$$

They represent the distributions of the r.v. x and y, resp., if the other variable ignored. Sections through distributions $f(x, y)$ are called conditional distributions. Given a fixed value $x = x_0$, the conditional p.d.f. of y is

$$p(y|x_0) = \frac{f(x_0,y)}{\int f(x_0,y)\, dy} = \frac{f(x_0,y)}{f_y(x_0)}. \tag{2.15}$$

From the marginal p.d.f $f_y(x)$ and the conditional density $p(y|x)$ the joint p.d.f. is given by

$$f(x,y) = p(y|x)\ f_y(x) \tag{2.16}$$

and the marginal p.d.f. $h(y)$ is given by

$$f_x(y) = \int_{-\infty}^{\infty} p(y|x) f_y(x)\, dx. \tag{2.17}$$

This equation is similar to the basic equation (1.01); the difference is in the normalization. While the distributions defined here are always normalized, the distributions appearing in equation (1.01) are for example cross sections and the integrals are true or measured total cross sections.

The definitions of expectation values and variances are straightforward generalized to multidimensional r.v.:

$$\begin{aligned} E(x) &= \iint x f(x,y)\, dx\, dy & V(x) &= \iint (x - E(x))^2 f(x,y)\, dx\, dy \\ E(y) &= \iint y f(x,y)\, dx\, dy & V(y) &= \iint (y - E(y))^2 f(x,y)\, dx\, dy. \end{aligned} \tag{2.18}$$

For a two dimensional p.d.f. there is an additional characteristic, the covariance

$$\sigma_{xy} = cov(x,y) = \iint (x - E(x))\,(y - E(y))\,f(x,y)\,dx\,dy. \tag{2.19}$$

The covariance can be expressed as $\sigma_{xy} = \rho_{xy}\sqrt{\sigma_x^2 + \sigma_y^2}$ with a correlation coefficient ρ_{xy}, which can take values between -1 and $+1$. The variables x and y are called uncorrelated, if $\rho_{xy} = 0$.

The p.d.f. of a random n-vector x, with components x_i, $i = 1 \ldots n$ may be written as $f(x)$. As a generalization of the variance one can define a covariance matrix by

$$V = V(x) = E\left((x - E(x))\,(x - E(x))^T\right), \tag{2.20}$$

which is a symmetric n-n matrix. The diagonal elements $V(x_i) = \sigma_i^2$ are called variances, the off-diagonal elements $V(x_i, x_j) = \sigma_{ij}$ are called covariances. When parameter values are quoted as the result of an experiment, usually only the square roots of the diagonal elements are given as parameter errors. The confidence level for multidimensional regions in parameter space depends however on the covariances of the parameters [**16**].

The n-dimensional normal distribution. The p.d.f. of the n-dimensional normal (or gaussian) distribution depends on the n means, and on the n^2 elements of the covariance matrix V (with $(n^2 + n)/2$ different elements):

$$f(x) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp(-\frac{1}{2}(x - \mu_x)^T\, V^{-1}\,(x - \mu_x)) \tag{2.21}$$

The n-dimensional gaussian distribution is the simplest model for the p.d.f. of n correlated random variables, since the only parameters are the n means and the $(n^2 + n)/2$ different elements of the covariance matrix.

The χ^2 distribution. If $x_1 \ldots x_n$ are independent variables, which all follow the normal distribution with mean 0 and variance 1, the sum u of the squares

$$u = \sum_{i=1}^{n} x_i^2 \tag{2.22}$$

follows the χ^2 distribution $\chi^2(n)$ with n degrees of freedom. The probability density is given by

$$f(u) = \frac{\frac{1}{2}(\frac{u}{2})^{n/2-1} e^{-u/2}}{\Gamma(\frac{n}{2})}. \tag{2.23}$$

The expectation value is n and the variance is $2n$. The χ^2 distribution is important for statistical tests; tables are given in standard textbooks on statistics. The 95 % confidence level of the χ^2 distribution with one degree of freedom for example is 3.84. If x is a single variable, distributed normally with mean 0 and standard deviation σ, a measured value $(\hat{x}/\sigma)^2$ will be ≤ 3.84 with 95 % probability. Thus a measured value $(\hat{x}/\sigma)^2 \leq 3.84$ is compatible with zero in the 95 % confidence limit.

Linear functions of random variables. A linear transformation $y = Bx$ of a random n-vector x with mean μ_x and covariance matrix $V(x)$ to a m-vector y is considered. The expectation value μ_y of y and the covariance matrix $V(y)$ of y are given by:

$$\mu_y = B\mu_x \quad \text{and} \quad V(y) = BV(x)B^T, \tag{2.24}$$

where B^T is the matrix transposed to the matrix B. The expression for $V(y)$ is the equation of standard error propagation.

2.3 Parameter estimation

The estimation of parameters from measured data is a standard problem in data analysis. The application of the maximum likelihood method for a certain class of problems is discussed below.

Assume, that the dependence of a cross section $f(x)$ on a variable x has been measured. There exists a model, which expresses $f(x)$ as a sum of m known functions $p_j(x)$ with coefficients a_j:

$$f(x) = \sum_{j=1}^{m} a_j p_j(x). \tag{2.25}$$

The m parameters (or coefficients) have to be determined from the data, which are given in form of histogram bin contents $\hat{f}_i, i = 1 \ldots n$. Each bin of width Δ is centered at a certain value x_i. Neglecting the variation of the functions $p_j(x)$ within a bin, the expected content of a histogram bin is

$$f_i = \Delta \cdot f(x) = \Delta \cdot \sum_{j=1}^{m} a_j p_j(x_i) = \sum_{j=1}^{m} A_{ij} a_j \qquad \text{with} \qquad A_{ij} = \Delta \cdot p_j(x_i). \tag{2.26}$$

The observed content of a histogram bin will follow a certain probability distribution, with a probability $P(\hat{f}_i|f_i)$ of observing $\hat{f}_i$, if the mean value is f_i. The product of all n probabilities,

$$L(a) = \prod_{i=1}^{n} P(\hat{f}_i|f_i) \tag{2.27}$$

is a function of the values of the parameters, and is called likelihood function. According to the maximum likelihood method [**16**], the best estimates of the parameters a are given by those values $\hat{a}$, for which the likelihood function takes on its largest value.

In applications of the maximum likelihood method usually a search is made for the minimum of the negative logarithm of the likelihood function:

$$S(a) = -\sum_{i-1}^{n} \ln P(\hat{f}_i|f_i). \tag{2.28}$$

Since the number of entries in a histogram bin will follow the Poisson distribution, the corresponding expression for $P(\hat{f}_i|f_i)$ can be inserted; dropping all constant terms, one gets

$$S(a) = \sum_{i=1}^{n} f_i - \sum_{i=1}^{n} \hat{f}_i \ln f_i. \tag{2.29}$$

Methods for the determination of the minimum usually are based on the approximation of $S(a)$ by a quadratic function, then the minimum can be determined by standard matrix methods. The derivatives of $S(a)$ w.r.t. the parameters at an approximate solution $\tilde{a}$ are given by:

$$\frac{\partial S}{\partial a_j} = \sum_{i=1}^{n} A_{ij} - \sum_{i=1}^{n} \hat{f}_i \frac{A_{ij}}{f_i} \qquad \frac{\partial^2 S}{\partial a_j \partial a_k} = \sum_{i=1}^{n} \hat{f}_i \frac{A_{ij} A_{ik}}{f_i^2} \tag{2.30}$$

with f_i calculated using the approximate solution $\tilde{a}$. In matrix notation the quadratic approximation can be written in the form

$$S(a) = S(\tilde{a}) - (a - \tilde{a})^T h + \frac{1}{2}(a - \tilde{a})^T H (a - \tilde{a}), \tag{2.31}$$

where h and H with elements

$$h_j = -\frac{\partial S}{\partial a_j} \qquad H_{jk} = \frac{\partial^2 S}{\partial a_j \partial a_k} \tag{2.32}$$

are the (negative) gradient and the Hessian of $S(a)$, respectively. The minimum of the quadratic approximation above is defined by the condition $\nabla S = 0$,

$$-h + H(a - \tilde{a}) = 0, \tag{2.33}$$

which is solved by

$$a_{app} = \tilde{a} + H^{-1} h. \tag{2.34}$$

Since the obtained result a_{app} is based on the approximation of the true function $S(a)$ at $\tilde{a}$, several iterations have to be performed; in each iteration the result of the previous iteration replaces $\tilde{a}$. Convergence may be assumed, if both the expected change of $S(a)$ in one iteration,

$$(\Delta S)_{exp} = -\frac{1}{2}(a_{app} - \tilde{a})^T h \tag{2.35}$$

and the actual change of $S(a)$ are small compared to 1. The method is stable, if each iteration includes a search for the minimum of the function

$$S(t) = S(\tilde{a} + t(a_{app} - \tilde{a})), \tag{2.36}$$

depending on one parameter t. The result of the minimum search,

$$a_{min} = \tilde{a} + t_{min}(a_{app} - \tilde{a}), \tag{2.37}$$

then replaces $\tilde{a}$ for the next iteration. It can be shown, that (negative) log likelihood functions are in fact approximately quadratic function, at least near the solution; therefore the value t_{min} is usually close to 1 and convergence with result $\hat{a}$ is reached within few iterations.

A simpler method for the solution of the problem is based on the approximation of the Poisson distribution by the gaussian distribution. Under the conditions mentioned

in chapter 2.1 the Poisson probability may be approximated by the value of the gaussian density with $\sigma_i^2 = f_i$. Using the additional approximation $\sigma_i^2 = \hat{f}_i$ the problem becomes easily solvable. Inserting for $P(\hat{f}_i|f_i)$ the expression for the gaussian density, $S(a)$ becomes

$$S(a) = \frac{1}{2}\sum_{i=1}^{n} \frac{(\hat{f}_i - f_i)^2}{\sigma_i^2} \qquad \sigma_i^2 = \hat{f}_i, \tag{2.38}$$

which is (except for the trivial factor 1/2) the expression to be minimized according to the least squares principle [17]. In this case the derivatives become

$$\frac{\partial S}{\partial a_j} = -\sum_{i=1}^{n} A_{ij}\frac{(\hat{f}_i - f_i)}{\sigma_i^2} \qquad \frac{\partial^2 S}{\partial a_j \partial a_k} = \sum_{i=1}^{n} \frac{A_{ij}A_{ik}}{\sigma_i^2} \tag{2.39}$$

In matrix notation the function $S(a)$ can be written in the form

$$S(a) = S(\tilde{a}) - (a - \tilde{a})^T h + \frac{1}{2}(a - \tilde{a})^T H (a - \tilde{a}), \tag{2.40}$$

identical to the Poisson case and therefore can be treated in the same way. However, since $S(a)$ in this case is really quadratically in a, no iteration is necessary and the result

$$\hat{a} = H^{-1}h \tag{2.41}$$

is obtained in one step, which does not require an approximate starting value $\tilde{a}$. Thus the least squares method may be used to calculate an approximate solution, required by the method based on the Poisson distribution.

Since $\hat{a}$ of equation above is a linear function of the gradient h, which itself is a linear function of the measured data $\hat{f}_i$, the simple formula of error propagation can be used to calculate the covariance matrix $V(\hat{a})$. In order to derive the result with matrix methods, a n-by-n weight matrix W is introduced, which is the inverse of the covariance matrix $V(\hat{f})$ of the measured data. Since the data are uncorrelated, the matrix $V(\hat{f})$ is a diagonal matrix with diagonal elements σ_i^2, and the weight matrix W is diagonal too, with diagonal elements $1/\sigma_i^2$. The (negative) gradient h and the Hessian H are given by the matrix expressions

$$h = A^T W \hat{f} \qquad H = A^T W A, \tag{2.42}$$

and the full solution can be written in the form

$$\hat{a} = H^{-1}h = (A^T W A)^{-1} A^T W \hat{f}. \tag{2.43}$$

Using the formula of error propagation, the covariance matrix $V(\hat{a})$ is

$$V(\hat{a}) = (A^T W A)^{-1} A^T W W^{-1} W A (A^T W A)^{-1} = (A^T W A)^{-1} = H^{-1}. \tag{2.44}$$

It can be shown that the same simple formula $V(\hat{a}) = H^{-1}$ applies to the result, obtained with the maximum likehood method based on the Poisson distribution. This is at least true for the asymptotic case of high statistics, but is a good approximation already for low statistics.

3. PARAMETRIZATION OF FUNCTIONS

3.1 INTERPOLATING SPLINE FUNCTIONS

Spline functions [18] are smooth interpolating functions. Besides applications in graphics, spline functions are increasingly used for numerical problems, for example in methods of solving boundary-value problems of differential equations. The standard application of spline functions is the interpolation between pairs (y_i, x_i) with $x_i \in [a, b]$. The set $Y = \{y_0, y_1 \dots y_n\}$ of $(n+1)$ real numbers represents values of a function $f(x)$ at abscissa values x_i, called knots. The set $X = \{x_0, x_1 \dots x_n\}$ with $x_0 = a$ and $x_n = b$ can be considered as a partition of the interval $[a, b]$. A cubic spline function $S(x)$ with $S(x_i) = y_i$, $i = 0 \dots n$, is a twice continuously differentiable function on $[a, b]$, coinciding on every subinterval $[x_i, x_{i+1}], i = 0 \dots (n-1)$ with a polynomial of third degree:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3. \tag{3.01}$$

At every inner knot x_i the two polynomials of the adjacent subintervals agree in the function values and in the values of the first two derivatives.

A cubic spline function with total $4n$ coefficients is not uniquely defined. The requirements $S(x_i) = y_i$ give $(n+1)$ conditions for the coefficients, and with the $(n-1)$ conditions at the inner knots on $S(x)$, $S'(x)$ and $S''(x)$, there are in total $(n+1) + 3(n-1) = 4n - 2$ conditions for the $4n$ coefficients. Thus two degrees of freedom are left, which have to be fixed by two additional conditions. Often used conditions are the following:

(I) $S''(x_0) = 0 \quad ; \quad S''(x_n) = 0$ (natural spline)

(II) $S'(x_0) = y'_0 \quad ; \quad S'(x_n) = y'_n$ (complete spline).

A natural requirement for an interpolating function is a certain degree of smoothness. Since the local curvature $f''(x)/(1 + f'^2(x))^{3/2}$ of a function $f(x)$ can be approximated for $|f'(x)| \ll 1$ by f'', the integral

$$\int_a^b [f''(x)]^2 \, dx \tag{3.02}$$

appears to be a reasonable quantitative measure of the smoothness of a function $f(x)$. This quantity will be called the (total) curvature of a function $f(x)$ in an interval $[a, b]$ in the following.

The natural spline (condition (I) above) is the smoothest function to interpolate given support points (y_i, x_i) in the sense of the curvature (3.02). First a proof is given for the statement:

$$0 \le \int_a^b |f''(x) - S''(x)|^2 \, dx = \int_a^b |f''(x)|^2 \, dx - \int_a^b |S''(x)|^2 \, dx \tag{3.03}$$

for cubic spline functions $S(x)$ under conditions (I) and (II), and twice continuously differentiable functions $f(x)$. The first integral of equation (3.03) can be rewritten in

the form

$$\int_a^b |f''(x)-S''(x)|^2\,dx = \int_a^b |f''(x)|^2\,dx - \int_a^b |S''(x)|^2\,dx - 2\int_a^b (f''(x)-S''(x))S''(x)\,dx \tag{3.04}$$

The last term is integrated by parts. For each subinterval $[x_{i-1}, x_i], i = 1 \ldots n$,

$$\begin{aligned}\int_{x_{i-1}}^{x_i} (f''(x) - S''(x))S''(x)dx &= (f'(x) - S'(x))S''(x)\Big|_{x_{i-1}}^{x_i} \\ &\quad - \int_{xi-1}^{x_i} (f'(x) - S'(x))S^{(3)}(x)\,dx \\ &= (f'(x) - S'(x))S''(x)\Big|_{x_{i-1}}^{x_i} - (f(x) - S(x))S^{(3)}(x)\Big|_{x_{i-1}}^{x_i} \\ &\quad + \int_{x_{i-1}}^{x_i} (f(x) - S(x))S^{(4)}(x)dx \end{aligned} \tag{3.05}$$

And adding up all terms one gets with $S^{(4)}(x) = 0$

$$\begin{aligned} 0 \le \int_a^b |f''(x) - S''(x)|^2\,dx & \\ &= \int_a^b |f''(x)|^2 dx - \int_a^b |S''(x)|^2 dx - 2(f'(x) - S'(x))S''(x)\Big|_a^b \\ &\quad + 2\sum_{i=1}^{n} (f(x) - S(x))S'''(x)\Big|_{x_{i-1}}^{x_i} \end{aligned} \tag{3.06}$$

The last term (the sum) vanishes because of the interpolation condition $S(x_i) = y_i$, $i = 0 \ldots n$, and the term before vanishes under conditions (I) and (II). Thus the inequality

$$\int_a^b |S''(x)|^2 dx \le \int_a^b |f''(x)|^2 dx, \tag{3.07}$$

follows and proves the statement (3.02). The inequality (3.07) implies that, among all twice continuously differentiable functions $f(x)$ with $f(x_i) = x_i$ the spline function $S(x)$ with condition (I) (natural spline) minimizes the total curvature (3.03). It also minimizes approximately the 'strain energy' in a curved elastic bar

$$\int_a^b \frac{(f''(x))^2}{(1 + f'(x))^{5/2}}\,dx \tag{3.08}$$

and this property is the origin of the name 'spline'. As an example the interpolation of 10 given points (y_i, x_i) by a spline function is shown in Figure 3, and is compared with the interpolation by a polynomial (of degree 9). The polynomial shows large osciallations near the endpoints, typical for the interpolation of equidistant data by a high order polynomial [**19**]. The advantages of the spline interpolation are apparent.

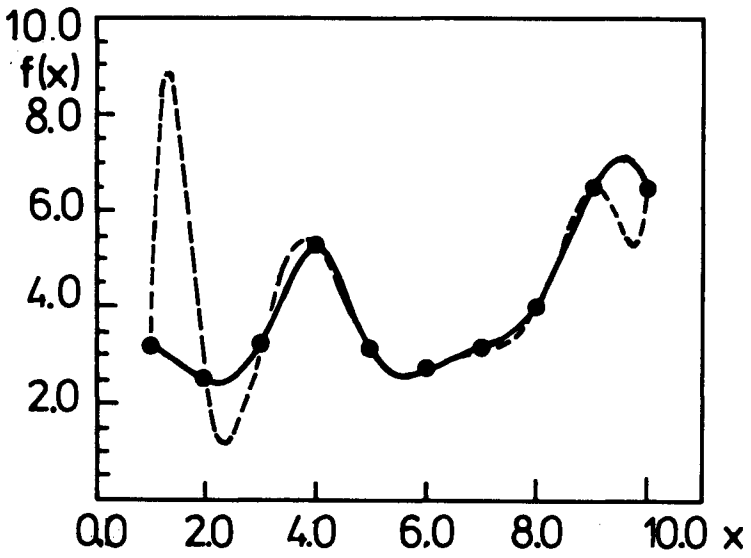

Figure 3. Interpolation of 10 given points (y_i, x_i) by a cubic spline function (full curve) and by a nineth order polynomial (dashed curve).

However, smoothness is not the only criterion for an interpolating function. A more important criterion is the accuracy of approximation of a given function $f(x)$ by $S(x)$. From an approximation point of view, the properties of spline functions with condition (I) are not optimal (unless really $f''(x_0) = f''(x_n) = 0$). If the true slopes at the end points are known, condition (II) gives a better approximation, and if the values of the second derivatives are known at the end points, the condition

(III) $S''(x_0) = y_0''$; $S''(x_n) = y_n''$

can be used. If nothing is known about end point derivatives, one can use as a general condition

(IV) $S'''(x)$ continuous across x_1 and x_{n-1} (not-a-knot condition).

The not-a-knot condition means, that the first and last inner knots are not active.

Spline functions $S(x)$ have optimal approximation properties (for a quantitative treatment see[18]). The difference $|f(x) - S(x)|$ is bounded by a quantity proportional to the fourth power of the knot spacing. Even the k-th derivatives of $f(x)$ for $k = 1, 2$ and 3 are well approximated (with a bound on the difference to the true derivative proportional to the $(4-k)$-th power of the knot spacing). The natural spline condition (I) introduces an error proportional to the square of the knot spacing near the ends, which is avoided by the not-a-knot condition (IV) and therefore this condition seems to be preferable as a general condition. In fact the not-a-knot condition has been used for the construction of the interpolating spline in Figure 3.

The determination of spline functions $S(x)$ for given support points (y_i, x_i) is a stable process for all conditions (I) - (IV); algorithms for the determination of the coefficients a_i, b_i, c_i, d_i, $i = 1 \ldots (n-1)$ can be found in [18], [20]. The construction of

higher order spline functions is possible, in practice however there is rarely a reason to go beyond cubic spline functions.

3.2 B-SPLINES

A representation of spline functions $S(x)$, different from the one given by equation (3.01), but equivalent, is provided by the so called basis splines or short B-splines [**20**]. A B-spline is itself a spline function. A single B-spline of order k is nonzero only in a limited range of x (basis). A spline function $S(x)$ of order k can be represented by s sum of B-splines $B_{j,k}(x)$ according to

$$S(x) = \sum_j a_j B_{j,k}(x), \tag{3.09}$$

where the functions $B_{j,k}(x)$ are B-splines of order k. The representation of spline functions by a linear combination of B-splines has numerical advantages in certain applications, for example least squares fits of a spline function to data become linear.

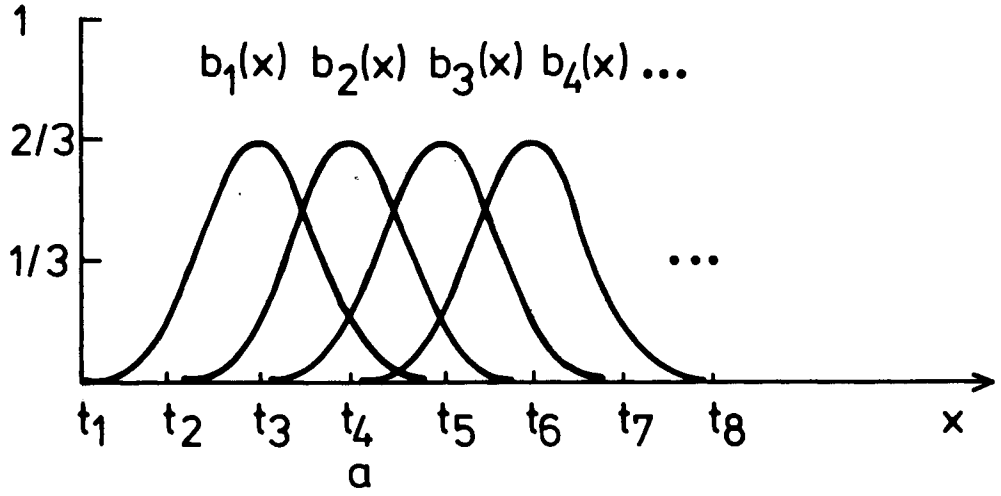

Figure 4. Sequence of cubic B-splines and equidistant knots.

B-splines are defined over a nondecreasing sequence $\{t_j\}$ of knots. Order $k = 1$ B-splines are defined by

$$B_{j,1} = \begin{cases} 1 & t_j \le x < t_{j+1} \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

Higher order B-splines $B_{j,k}(x)$ (with $k > 1$) are defined by a recurrence relation, which allows to compute $B_{j,k}(x)$ by positive linear combinations of positive quantities:

$$B_{j,k} = \frac{x - t_j}{t_{j+k-1} - t_j} B_{j,k-1}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+1}} B_{j+1,k-1}(x). \tag{3.11}$$

B-splines $B_{j,k}(x)$ have the property

$$B_{j,k}(x) = \begin{cases} > 0 & t_j < x < t_{j+k} \\ 0 & \text{otherwise} \end{cases}, \tag{3.12}$$

i.e. they take on only positive values (or zero). In any particular interval $[t_j, t_{j+1}]$ only the k B-splines $B_{j-k+1,k}(x) \cdots B_{j,k}(x)$ may be nonzero. B-splines, as defined in equations (3.10) and (3.11), are normalized in the sense, that at any particular x-value their sum is equal to 1:

$$\sum_j B_{j,k}(x) = \sum_{j=l+1-k}^{l} B_{j,k}(x) = 1 \qquad t_l \le x < t_{l+1}. \tag{3.13}$$

In the following only cubic B-splines ($k = 4$) are considered. For the case of B-splines $B_{j,4}(x)$ with equidistant knots, which are often sufficient, explicit formulas are given, denoting these special B-splines by $b_j(x)$. If m B-splines $b_j(x)$ are used for the parametrization of a function for $a \le x \le b$, the distance between adjacent knots is $d = (b-a)/(m-3)$. In total there are $(m+4)$ knots, with $t_4 = a$, $t_{m+1} = b$ and in general $t_j = a + (j-4)d$ (see Figure 4). The explicit formulas for the B-splines $b_j(x)$ in terms of a variable z with $0 \le z < 1$ are:

$$b_j(x) = \begin{cases} \frac{1}{6}z^3 & z = (x - t_j)/d & t_j \le x < t_{j+1} \\ \frac{1}{6}[1 + 3(1 + z(1-z))z] & z = (x - t_{j+1})/d & t_{j+1} \le x < t_{j+2} \\ \frac{1}{6}[1 + 3(1 + z(1-z))(1-z)] & z = (x - t_{j+2})/d & t_{j+2} \le x < t_{j+3} \\ \frac{1}{6}(1-z)^3 & z = (x - t_{j+3})/d & t_{j+3} \le x < t_{j+4} \\ 0 & \text{otherwise} & \end{cases} \tag{3.14}$$

A single B-spline $b_j(x)$ together with the derivatives is shown in Figure 5.

The m coefficients for a spline function in the parametrization (3.09) with B-splines $b_j(x)$, interpolating $(m-2)$ equidistant data points (y_i, x_i) with $x_i = t_i$, $i = 4, \ldots m+1$, are determined by a set of linear equations. The first $(m-2)$ equations

$$6\,y_j = a_{j-3} + 4a_{j-2} + a_{j-1} \qquad j = 4 \ldots m+1 \tag{3.15}$$

are given by the interpolation conditions. The additional two conditions given by the not-a-knot condition (see chapter 3.1) are

$$\begin{aligned} -a_1 + 4a_2 - 6a_3 + 4a_4 - a_5 &= 0 \\ -a_{m-4} + 4a_{m-3} - 6a_{m-2} + 4a_{m-1} - a_m &= 0. \end{aligned} \tag{3.16}$$

In a parametrization (3.09) by B-splines, each function value $S(x)$ at a certain x as well as derivatives $S'(x) \ldots$ and the integral of $S(x)$ over a x-region is a linear combination of the coefficients a_j. This property allows linear least squares fits to data (y_i, x_i) of spline functions in the B-spline parametrization. Having determined the coefficients a_j together with their covariance matrix, error propagation for interpolated function values, derivatives etc. is straightforward.

3.3 ORTHOGONAL FUNCTIONS

The concept of orthogonal functions is of particular importance in many fields of numerical and statistical analysis. A system $\{p_j(x)\}$ of functions $p_j(x)$, defined for $a \le x \le b$, is called orthogonal, if the inner product of each two functions (p_j, p_k)

$$(p_j, p_k) = \int_a^b p_j(x) p_k(x)\, dx = 0 \qquad \text{for} \quad j \ne k. \tag{3.17}$$

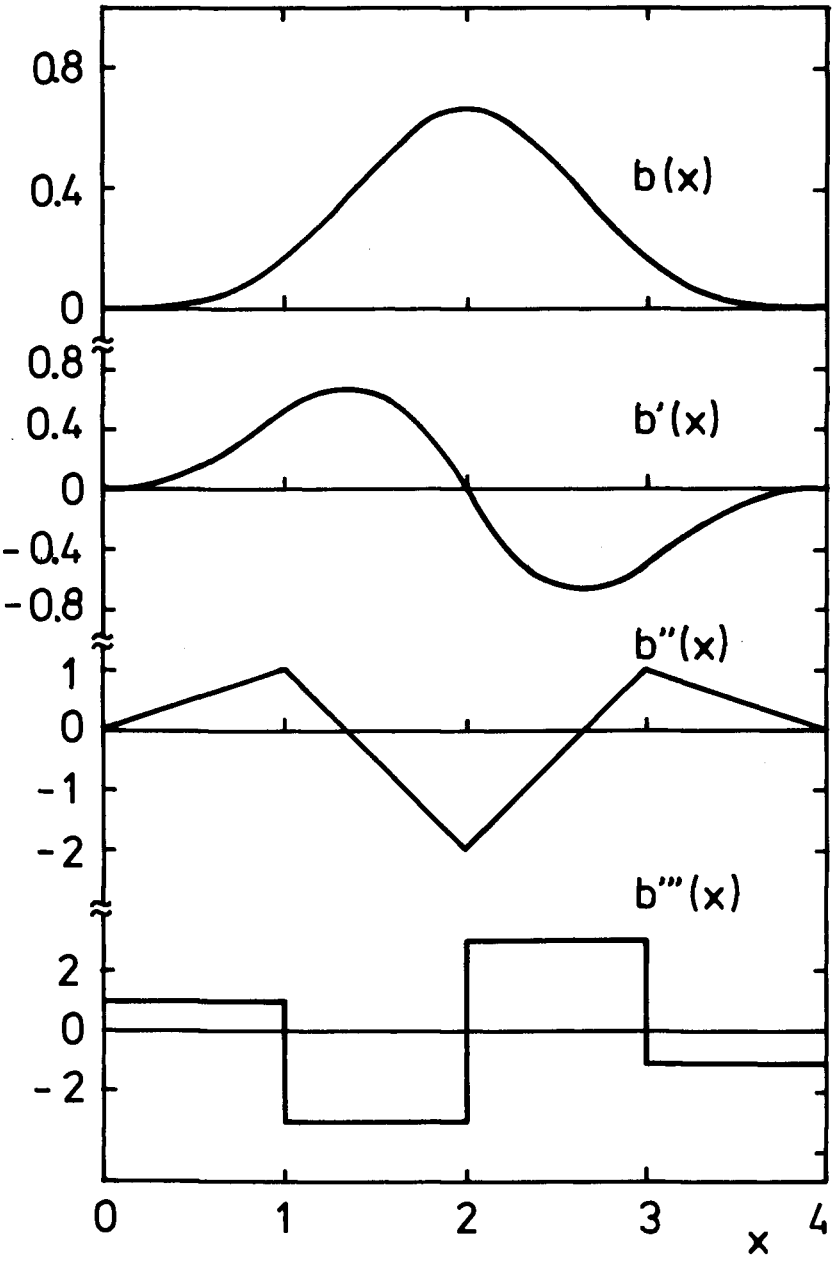

Figure 5. Single B-spline and derivatives, for equidistant knots with distance $d = 1$ between adjacent knots.

Orthogonal functions are called normalized, if the values N_j, defined by

$$\int_a^b p_j^2(x)\,dx = N_j \tag{3.18}$$

are equal to 1 for all values of j. By an orthogonalization procedure a system of orthogonal functions can be constructed from a system of linear independent functions [21]. A piecewise continuous function $g(x)$ can be expanded in terms of a system of normalized

orthogonal functions,

$$g(x) = \sum_{j=1}^{\infty} a_j p_j(x) \qquad \text{with} \qquad a_j = \int_a^b p_j(x) g(x) dx\,, \tag{3.19}$$

where the factors a_j are called expansion coefficients or components of the function $f(x)$ with respect to the system $\{p_j(x)\}$. The deviation between the function $f(x)$ and a finite sum with n terms can be measured by the quadratic expression

$$M_m = \int_a^b \left[f(x) - \sum_{j=1}^{m} a_j p_j(x) \right]^2 dx \tag{3.20}$$

According to the principle of least squares the quantity M_m should be as small as possible. It can be shown [**21**], that

$$\sum_{j=1}^{m} a_j^2 \leq \int_a^b [f(x)]^2 \, dx \tag{3.21}$$

A system of orthogonal functions is called complete, if for any positive number ϵ there is an index m, such that $M_m < \epsilon$. In practice usually only a small number of terms is necessary, since the magnitude of $|a_j|$ falls rapidly with increasing index j above a certain index. This feature allows an efficient representation of functions $f(x)$ by finite sums, which can be evaluated very fast, using recurrence relations for the evaluation of the $p_j(x)$ [**19**].

An important example of a complete orthogonal system of functions is given by the functions $1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots$, which is orthogonal in the interval $0 \leq x \leq 2\pi$. A periodic function, with period 2π, can be expanded in a Fourier sum

$$f(x) = \frac{a_0}{2} + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \sin \nu x), \tag{3.22}$$

where the expansion coefficients a_ν and b_ν (Fourier coefficients) are given by the formulas:

$$a_\nu = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos \nu x \, dx \qquad b_\nu = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin \nu x \, dx. \tag{3.23}$$

It can be shown, that the trigonometric interpolation (3.22) converges to the given function $f(x)$ at every point on the range [**22**]. The asymptotic behaviour of the Fourier coefficients depends on the degree of differentiability of $f(x)$:

$$\sqrt{a_\nu^2 + b_\nu^2} = O\left(\frac{1}{|\nu|^{r+1}} \right) \tag{3.24}$$

for a 2π-periodic function $f(x)$ having an absolutely continuous r-th derivative [**18**]. Another interesting property is the fact, that each term of the expansion represents an independent contribution to the total curvature:

$$\int_o^{2\pi} [f''(x)]^2 = \pi \sum_{\nu=1}^{\infty} \nu^4 (a_\nu^2 + b_\nu^2). \tag{3.25}$$

Equations (3.24) and (3.25) show, that for a smooth function $f(x)$ (for example $r \geq 3$ in equation (3.24)) the coefficients of the higher terms of the expansion, which have large contributions to the curvature, fall rapidly with increasing index ν. One method to smooth a set of empirical data (y_i, x_i) with statistical fluctuations is the determination of the Fourier coefficients (analysis), the attenuation of the coefficients with higher index representing noise only, and the reconstruction (synthesis) according to equation (3.22) [23].

A system of orthogonal polynomials can be constructed from the monomials 1, x, $x^2 \ldots$. For the region $-1 \leq x \leq 1$ the orthogonal polynomials are identical to the Legendre polynomials (apart from constant factors). Orthogonal polynomials are often used to approximate an empirical function given by a discrete set of points (y_i, x_i), $i = 1 \ldots n$, if no specific parametrization is known. Following the least squares principle, orthogonal polynomials for the discrete set can be constructed from a recurrence relation, starting from a constant and a linear term, by the requirement

$$\sum_{i=1}^{n} w_i p_j(x_i) p_k(x_i) = \delta_{jk}, \tag{3.26}$$

where w_i is the weight of an individual data point, which can be defined by $w_i = 1/\sigma_i^2$ for the standard deviation σ_i [17]. For the normalized orthogonal polynomials $p_j(x)$ the coefficients a_j of the approximation

$$f(x) = \sum_j a_j p_j(x) \tag{3.27}$$

can be calculated by summation:

$$a_j = \sum_{i=1}^{n} w_i p_j(x_i)\, y_i. \tag{3.28}$$

Because of the orthogonality of the functions $p_j(x)$, the covariance matrix $V(a)$ of the coefficients is a unit matrix I (this corresponds to a least squares fit with $H = I$, compare chapter 2.3). This property allows statistical χ^2 tests on the significance of each coefficient a_j. If all coefficients a_j for $j > m_0$ are compatible with zero, the discrete set (y_i, x_i) can be approximated by m_0 terms of the expansion (3.27), the lowest order polynomial consistent with the data.

4. UNFOLDING OF CONTINUOUS DISTRIBUTIONS

4.1 UNFOLDING OF PERIODIC FUNCTIONS

In this chapter the difficulties inherent in unfolding procedures are discussed in a special case, which makes them clearly apparent. Consider the case of a function $f(x)$ in the range $0 \leq x \leq 2\pi$, periodic with the period 2π, which is measured with a gaussian resolution function.

Using the formulae (3.23) for the determination of the Fourier coefficients, a piecewise continuous function $f(x)$ can be expanded according to

$$f(x) = \frac{a_0}{2} + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \sin \nu x) \tag{4.01}$$

with $a_\nu, b_\nu \to 0$ for $\nu \to \infty$. The folding by a gaussian resolution function with a standard deviation σ gives

$$g(y) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) f(x)\, dx. \tag{4.02}$$

For a single term $\cos \nu x$ of the expansion (4.01) one gets

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \cos \nu x\, dx = \exp\left(-\frac{\nu^2\sigma^2}{2}\right) \cos \nu y \tag{4.03}$$

This means, that a single term in the expansion (4.01) has, after folding, the same form as before. The amplitude however is attenuated by a factor $\exp(-\nu^2\sigma^2/2)$, as shown in Figure 6 for two terms of the expansion (4.01).

This interesting result which is true for all $\cos \nu x$ and $\sin \nu x$ terms, shows how to unfold a measured periodic function. The measured function has to be expanded in the form

$$g(y) = \frac{\alpha_0}{2} + \sum_{\nu=1}^{\infty} (\alpha_\nu \cos \nu x + \beta_\nu \sin \nu x) \tag{4.04}$$

by formulae which are equivalent to formulae (3.23). Unfolding and reconstruction of the original function $f(x)$ is then done by

$$a_\nu = \exp\left(\frac{\nu^2\sigma^2}{2}\right) \alpha_\nu \qquad b_\nu = \exp\left(\frac{\nu^2\sigma^2}{2}\right) \beta_\nu \tag{4.05}$$

These exact formulae show very clearly the difficulties of unfolding. The coefficients α_ν and β_ν can only be determined with some statistical errors, while the true values become smaller with increasing value of ν; the multiplication in formulae (4.05) then means the multiplication of the statistical errors with a rapidly increasing exponential factor, and the unfolded result would soon be dominated by statistical fluctuations. The reconstruction of the coefficients a_ν, b_ν above a certain value of the index ν becomes meaningless. This means that one either obtains very large unwanted fluctuations in

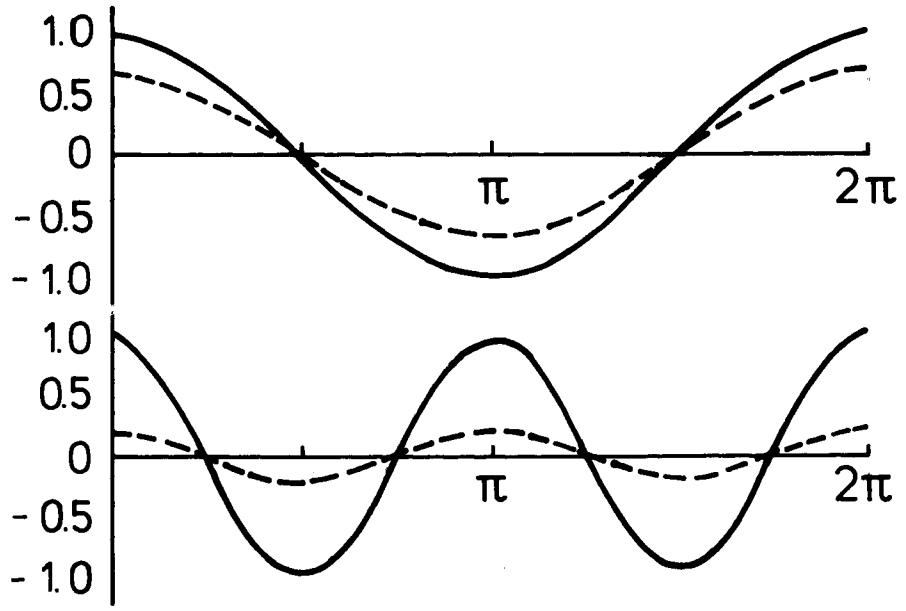

Figure 6. Graph of the functions $\cos x$ and $\cos 2x$ (full curves) and of the same functions after folding with a gaussian resolution function with $\sigma^2 = 3/4$.

the solution, or one has to limit the number of Fourier terms which prevents the finer structures to be resolved. The latter is of course expected as a consequence of the limited resolution.

4.2 Discretization

The fundamental equation (1.01) relates the true distribution $f(x)$ to the distribution $g(y)$ measured in an experiment. In actual experiments the measured distribution usually contains some background contribution from other sources. It is assumed, that the background $b(y)$ can be either measured independently or calculated; in any case, the background is assumed to be known. Thus, for a given true distribution $f(x)$, the expected distribution $g(y)$ in the measured variables y can be written in the form:

$$g(y) = \int_a^b A(y,x) f(x) dx + b(y). \tag{4.06}$$

The distribution $\hat{g}(y)$ actually measured differs from the expected distribution $g(y)$ by some statistical errors $\epsilon(x)$. To obtain the true dependence $f(x)$ from the data, equation (4.06) has to be discretized, representing the continuous function $f(x)$ by a finite set of coefficients $a_1, a_2 \ldots a_j \ldots a_m$. The discretization, as described below, results in an equation of the form

$$g = Aa + b, \tag{4.07}$$

where g, a and b are vectors and A is a matrix, representing the response function $A(y,x)$. As remarked in chapter 1, acceptance and resolution in high-energy physics

experiments are usually defined only implicitly by MC-procedures and this fact has to be taken into account in the discretization method.

The discretization of equation (4.06) is done in two steps. In a first step the function $f(x)$ is parametrized by a sum

$$f(x) = \sum_{j=1}^{m} a_j p_j(x), \tag{4.08}$$

using a certain set of basis functions $p_j(x)$ to be specified later. The parametrization (4.08) allows to perform the integration:

$$\int_a^b A(y,x) f(x) dx = \sum_{j=1}^{m} a_j \left[\int_a^b A(y,x) p_j(x) dx \right] = \sum_{j=1}^{m} a_j A_j(y)$$
$$\text{with} \quad A_j(y) = \int_a^b A(y,x) p_j(x) dx. \tag{4.09}$$

Now equation (4.06) can be rewritten in the form

$$g(y) = \sum_{j=1}^{m} a_j A_j(y) + b(y). \tag{4.10}$$

The expected distribution $g(y)$ is expressed by a superposition of functions $A_j(y)$, each representing one term $p_j(x)$ in the representation (4.08).

The second discretization step is the representation of all y-dependent functions in equation (4.10) by histograms, assuming a certain set of bin-limits $y_0, y_1 \ldots y_n$:

$$g_i = \int_{y_{i-1}}^{y_i} g(y) dy \qquad A_{ij} = \int_{y_{i-1}}^{y_i} A_j(y) dy \qquad b_j = \int_{y_{i-1}}^{y_i} b(y) dy. \tag{4.11}$$

Using this discretization equation (4.06) can be written in the form of equation (4.07). g and b are n-vectors, representing histograms of the measured quantity y. The vector a is a m-vector of coefficients a_j, and A is a n-by-m matrix of elements A_{ij}; column A_j of matrix A represents the histogram in y for $f(x) = p_j(x)$. The elements A_{ij} of matrix A are defined by the MC-events. Each MC-event, with true value x, is added to histogram $A_j(y)$ with a weight proportional to $p_j(x)$. In order to avoid negative weights and to simplify a proper normalization, the conditions

$$p_j(x) \geq 0 \qquad \sum_{j=1}^{m} p_j(x) \equiv 1 \tag{4.12}$$

are required. That is, the sum of all weights $p_j(x)$ for a given MC-event is equal to 1. In addition an overall weight for the MC-events has to be defined, such that the resulting distribution $f(x)$ is correctly normalized. In particle reactions, where $f(x)$ is a cross section, the ratio of event number to cross section is called integrated luminosity $\int L\, dt$:

$$N(x) = f(x) \int L\, dt. \tag{4.13}$$

The integrated luminosity of the experiment has of course to be known, either from the experimental conditions or by the measurement of a monitor reaction. For the MC event simulation one can define a luminosity as well, by

$$\int L_{MC}\, dt = \frac{N_{MC}}{\int f_0(x)\, dx}, \tag{4.14}$$

if the MC events are generated according to $f_0(x)$. The simplest way is to use an unit cross section $f_0(x) \equiv 1$. If in this case the ratio of the integrated luminosities

$$\frac{\int L_{exp}\, dt}{\int L_{MC}\, dt} \tag{4.15}$$

is used as an overall weight of the MC events, the resulting $f(x)$ will be directly the correctly normalized cross section.

Some remarks are in order concerning the discretization. The choice of the basis function $p_j(x)$ is an important point. The simplest choice compatible with the conditions of eq.(4.12) is :

$$p_j(x) = \begin{cases} 1 & \text{for } \; t_{j-1} \leq x < t_j \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

with a set of knots $t_0, t_1, t_2 \ldots t_m$. This choice has the property, that the coefficients a_j directly represent a histogram of the solution $f(x)$. However, the function $f(x)$ would have discontinuities, and the step function approximation of the integrand in eqation (4.09) is inaccurate. An obvious choice for the basis functions $p_j(x)$ are cubic B-splines $b_j(x)$, which satisfy the requirements of eqation (4.12) and provide an accurate integration. Note, that the step functions defined in eqation (4.16) are also B-splines, of order $k = 0$, while cubic B-splines are of order $k = 4$. Besides the accurate integration cubic B-splines have further advantages, to be discussed later. Using cubic B-splines, the solution $f(x)$ is a smooth curve, determined by the m coefficients a_j. The final result should of course be represented by data points with error bars. These data points can be obtained by integration of $f(x)$ over small regions in x,

$$f_k = \left(\int_{x_{k-1}}^{x_k} f(x) dx \right) / (x_k - x_{k-1}) = \left(\sum_{j=1}^{m} a_j \int_{x_{k-1}}^{x_k} p_j(x) dx \right) / (x_k - x_{k-1}), \tag{4.17}$$

with the values f_k representing average values of the function $f(x)$ in $[x_{k-1}, x_k]$. Since the values f_k are linear functions of the coefficients a_j, the calculation of errors is straightforward.

In general it will be more efficient to generate MC event according to a distribution $f_0(x)$, which is close to the true distribution $f(x)$. The same definition of the integrated MC luminosity may be used as before, but the resulting $f(x)$ has to be multiplied by $f_0(x)$.

4.3 UNFOLDING WITHOUT REGULARIZATION

The discretization derived in the preceeding chapter facilitates unfolding through a fit of the linear expression $g = Aa + b$ to the data $\hat{g}$. The problem belongs to the class of

problems discussed in chapter 2.3. The solution $\hat{a}$ obtained in the unfolding case usually looks unsatisfactory: the resulting distribution

$$\hat{f}(x) = \sum_{j=1}^{m} \hat{a}_j p_j(x) \tag{4.17}$$

can show an oscillating behaviour, as mentioned in chapter 1 and discussed quantitatively in chapter 4.1, with fluctuations much larger than any physically motivated expectations.

In order to analyse the reason for this unwanted behaviour in detail, the solution of the problem is derived in a different way. One iteration step of the algorithm of chapter 2.3 is considered, which consists in the determination of the minimum of the quadratic approximation of the negative log likelihood function, starting from a previous estimate $\tilde{a}$:

$$S(a) = S(\tilde{a}) - (a - \tilde{a})^T h + \frac{1}{2}(a - \tilde{a})^T H (a - \tilde{a}). \tag{4.18}$$

In the following the parameters are transformed to a new basis. Since the matrix H is symmetric, it can be transformed to a diagonal matrix D,

$$D = U_1^T H U_1, \tag{4.19}$$

where the matrix D contains in its diagonal the real eigenvalues D_{jj} of H, which are positive because the matrix H is positive definite; the matrix U_1 is an orthogonal matrix with the property $U_1^T U_1 = I_{mm}$ and contains the eigenvectors u_j corresponding to the eigenvalues in its columns. The eigenvalues may be arranged in decreasing order $D_{11} \geq D_{22} \geq \ldots \geq D_{mm}$; in typical applications they decrease by several orders of magnitudes. A diagonal matrix $D^{1/2}$ with the property $D^{1/2} D^{1/2} = D$ can be defined, which has the positive square roots of D_{jj} in the diagonal. A transformation is defined between the parameter vector a and another vector a_1 by

$$a = U_1 D^{-1/2} a_1. \tag{4.20}$$

Inserting this expression into equation (4.18) one gets after omitting constant terms,

$$S(a_1) = -a_1^T D^{-1/2} U_1^T (H\tilde{a} + h) + \frac{1}{2} a_1^T a_1. \tag{4.21}$$

From the minimum condition $\nabla S = 0$ the solution

$$\hat{a}_1 = D^{-1/2} U_1^T (H\tilde{a} + h) \tag{4.22}$$

is obtained directly. The remarkable feature of the parameters $\hat{a}_1$, achieved by the transformation, is, that the covariance matrix $V(\hat{a}_1)$ is equal to the unit matrix I. This means, that the different components $(\hat{a}_1)_j$ of $\hat{a}_1$ are statistically independent and have a variance of 1. The result obtained is of course equivalent to the solution derived in chapter 2.4, which can be shown by a transformation back to a:

$$\hat{a} = U_1 D^{-1/2} \hat{a}_1 = U_1 D^{-1/2} D^{-1/2} U_1^T (H\tilde{a} + h) = H^{-1} h. \tag{4.23}$$

This equation also shows, that the solution vector can be expressed as a linear combination of the eigenvectors u_j (these are the columns of the matrix U_1).

The statistical independence of the components of $\hat{a}_1$ allows to test the statistical significance of every component independently. If the true value of a certain component is zero (or small compared to 1), the measured value $(\hat{a}_1)_j$ will follow a standard gaussian distribution, and $(\hat{a}_1)_j^2$ will follow a χ_1^2 distribution. Using a confidence level of 95 %, one can consider the j-th component to be compatible with zero, if $(\hat{a}_1)_j^2 \leq 3.84$. If all values $(\hat{a}_1)_j$ with $j > m_0$ are compatible with zero, they can be ignored, and the result can be expressed as a linear combination of the first m_0 eigenvectors. In fact it turns out that those insignificant components are the ones which cause the fluctuations in the full solution. This is seen easily, if the equation (4.20) is rewritten to the form

$$\hat{a} = \sum_{j=1}^{m} \left(\frac{1}{D_{jj}}\right)^{1/2} (\hat{a}_1)_j \, u_j. \tag{4.24}$$

Because of the factor $(1/D_{jj})^{1/2}$ (and the unit variance of $(\hat{a}_1)_j$) the insignificant components get a large weight factor in the full solution.

A sharp cut-off in the amplitudes at a certain index m_0 however also introduces some fluctuations in the solution, known as 'Gibbs phenomenon' in the theory of Fourier analysis and reconstruction of periodic functions. A smooth cut-off reducing these oscillations is provided by the regularization method, to be discussed in the next chapter.

4.4 UNFOLDING WITH REGULARIZATION

As explained in the preceeding chapter, unfolding by a straightforward fit without a cut-off will produce a fluctuating result. Mathematically the fluctuations are caused by insignificant components of the solution with their strong oscillations, which get a large weight in the unfolding. The magnitude of the fluctuations can be measured by several quantitities; one possible measure is the total curvature, introduced quantitatively in equation (3.02):

$$r(a) = \int [f''(x)]^2 \, dx. \tag{4.25}$$

One can also consider other measures of the smoothness of the solution, for example based on the square of the first derivative of $f(x)$ [**12**]. In this paper the discussion is restricted to the measure defined by equation (4.25). This quantity will take on large values for a strongly fluctuating solution, often orders of magnitudes larger than physically motivated expectations. This expectation of a smooth solution, an implicit a priori knowledge, can be used in the unfolding in the following way: a new function $R(a)$ is introduced by adding to the negative log likelihood function the total curvature $r(a)$, weighted by a factor τ:

$$R(a) = S(a) + \frac{1}{2}\tau \cdot r(a) \tag{4.26}$$

(the factor $\frac{1}{2}$ is introduced for later convenience). This method is called regularization method and the factor τ is called regularization parameter. If $f(x)$ is parametrized by a

sum of B-splines of order 4, the choice of equation (4.25) has the particular advantage, that $r(a)$ can be represented by a quadratic expression

$$r(a) = a^T C a \tag{4.27}$$

with a (constant) symmetric, positive semidefinite matrix C; such a regularization term is easily accounted for in the minimization. Apart from some factor, which can be absorbed in the regularization parameter τ, the matrix C has the form

$$C = \begin{pmatrix} 2 & -3 & 0 & 1 & 0 & 0 & \dots \\ -3 & 8 & -6 & 0 & 1 & 0 & \\ 0 & -6 & 14 & -9 & 0 & 1 & \\ 1 & 0 & -9 & 16 & -9 & 0 & \\ 0 & 1 & 0 & -9 & 16 & -9 & \\ 0 & 0 & 1 & 0 & -9 & 16 & \\ \vdots & & & & & & \ddots \end{pmatrix}$$

for cubic B-splines with equidistant knots.

Obviously, regularization terms can introduce a bias to the solution, which depends on the magnitude of the regularization parameter τ. For $\tau \to 0$ the effect of the regularization will vanish, and for $\tau \to \infty$ the result will become a linear function for $r(a)$ defined by equation (4.25). As is shown below, the magnitude of τ can be defined such that the bias will be negligible small compared to statistical errors; in effect, the regularization provides a smooth cut-off of higher order terms in the solution.

Using the transformation (4.20) from a to a_1, already defined in chapter 4.3, the espression (4.26) to be minimized can be rewritten in the form

$$R(a_1) = -a_1^T D^{-1/2} U_1^T (H\tilde{a} + h) + \frac{1}{2} a_1^T a_1 + \frac{1}{2}\tau \cdot a_1^T D^{-1/2} U_1^T C U_1 D^{-1/2} a_1 \tag{4.28}$$

The regularization term can be written as

$$\frac{1}{2}\tau \cdot a_1^T C_1 a_1 \quad \text{with} \quad C_1 = D^{-1/2} U_1^T C U_1 D^{-1/2}. \tag{4.29}$$

The regularized solution can be calculated using one more transformation of the parameters. The matrix C_1 is transformed to a diagonal matrix S by

$$S = U_2^T C_1 U_2, \tag{4.30}$$

where $U_2^T U_2^T = I_{mm}$; the eigenvalues S_{jj} can be arranged in increasing order $S_{11} \leq S_{22} \leq \dots \leq S_{mm}$. The additional transformation is defined by

$$a_1 = U_2 a' \tag{4.31}$$

and is a pure rotation in parameter space. Note that because of the pure rotation $a_1^T a_1 = a'^T a'$. The rotation yields

$$\frac{1}{2}\tau \cdot a'^T S a' \tag{4.32}$$

for the regularization term and the function to be minimized becomes

$$S(a') = -a'^T U_2^T D^{-1/2} U_1^T (H\tilde{a} + h) + \frac{1}{2} a'^T (I + \tau \cdot S) a'. \tag{4.33}$$

The regularized solution derived from the condition $\nabla S = 0$ is given by

$$\hat{a}' = (I + \tau \cdot S)^{-1} U_2^T D^{-1/2} U_1^T (H\tilde{a} + h)\,, \tag{4.34}$$

whereas the unregularized solution ($\tau = 0$), denoted by a bar, reads

$$\bar{a}' = U_2^T D^{-1/2} U_1^T (H\tilde{a} + h). \tag{4.35}$$

Due to the orthogonality of the rotation matrix U_2 the covariance matrix $V(\bar{a}')$ is still a unit matrix. The result can be transformed back to the coefficients of the basis functions $p_j(x)$ by

$$\bar{a} = U_1 D^{-1/2} U_2 \bar{a}', \tag{4.36}$$

yielding

$$\bar{f}(x) = \sum_{j=1}^{m} \bar{a}_j p_j(x). \tag{4.37}$$

An equivalent point of view is to consider the transformed set of basis functions $p'_j(x)$, and to write the result in the form

$$\bar{f}(x) = \sum_{j=1}^{m} \bar{a}'_j p'_j(x), \tag{4.38}$$

where the functions $p'_j(x)$ are linear combinations of the basis functions $p_j(x)$ and are orthogonal and normalized functions[1]. In this parametrization the curvature (4.25) is given by

$$\int \left[\bar{f}''(x)\right]^2 dx = \sum_{j=1}^{m} (\bar{a}'_j)^2 S_{jj}. \tag{4.39}$$

A normalized orthogonal function $p'_j(x)$ usually has $(j-1)$ zeros in the range, and since the eigenvalues S_{jj} are sorted in increasing order, the contribution to the curvature rises rapidly with increasing index j. The analysis of the values of the coefficients $\bar{a}'$ will show, that for $j > m_0$ they are compatible with zero within their statistical errors of 1. With a sharp cut-off at $j = m_0$, the result is expressed by m_0 statistically independent contributions and can be converted to just m_0 data points.

Now the effect of the regularization is considered. Equations (4.34) and (4.35) show, that the coefficients of the regularized solution are

$$\hat{a}'_j = \frac{1}{1 + \tau S_{jj}}\, \bar{a}'_j. \tag{4.40}$$

[1] The linear combinations are determined by the combined effect of the transformations (4.20) and (4.31).

Thus the coefficients of the regularized solution are obtained by the multiplication of the coefficients of the unregularized solution by a factor, which is close to 1 for all indices j with $S_{jj} \ll \tau^{-1}$. Since the values of S_{jj} increase rapidly, the attenuation factor will approach zero for $S_{jj} \gg \tau^{-1}$ after a transition region, where $S_{jj} \approx \tau^{-1}$. Regularization thus means a smooth cut-off, avoiding the already mentioned 'Gibbs phenomenon'. The sum of all factors can be considered as the effective number m_0 of independent contributions to the solution; for a given number m_0 the regularization parameter τ can be defined by

$$m_0 = \sum_{j=1}^{m} \frac{1}{1+\tau S_{jj}}. \tag{4.41}$$

The parameter m_0 has to be large enough, such that no significant coeffients are attenuated too much. A lower limit for m_0 can be obtained from statistical tests on the significance of the coefficients $\bar{a}'_j$. In typical applications the value of m_0 will be chosen just above the lower limit.

Having fixed the value of the regularization parameter τ, the regularized solution can be calculated using equation (4.40). Since the covariance matrix of $\bar{a}'$ is the unit matrix and the regularized solution is $\hat{a}' = (I+\tau \cdot S)^{-1}\bar{a}'$, the covariance matrix of $\hat{a}'$ is

$$V(\hat{a}') = (I+\tau \cdot S)^{-2}. \tag{4.42}$$

The result can then be transformed back by the transformations defined in equations (4.20) and (4.31).

The resulting coefficients have to be converted finally to a set of m_0 data points[1] $\hat{f}_k$, by integration of $\hat{f}(x)$ over small regions of x according to equation (4.17). The choice of these regions has of course consequences for the correlations between the data points, which should be as small as possible. One method to achieve this is the following. Since the function $p'_{m_0+1}(x)$ has just m_0 zeros, it seems optimal to define the m_0 regions around the zeros, with the (m_0-1) x-values, where the function $p'_{m_0+1}(x)$ has extreme values, taken as limits. This has the effect of suppressing the contribution of the term $\hat{a}'_{m_0+1}p'_{m_0+1}(x)$, which is attenuated by a factor of roughly 1/2, and takes into account the statistical precision and the resolution as a function of x. Each average value is expressed by a linear combination of the coefficients $\hat{a}'_j$ or $\hat{a}_j$, and error propagation is straightforward. Using the unfolding result $\hat{f}(x)$ as a weighting factor for the MC-events, the quality of the description of the measured distributions can be tested. This test can be extended to measured variables not directly used in the unfolding fit.

Numerical example of unfolding with regularization. In this example the Monte Carlo technique is used for the simulation of a measurement with limited acceptance and limited resolution. The measurement of a variable x with $0 \leq x \leq 2$ is simulated with the following properties of the measurement: The acceptance probability is assumed to be

$$P_{acc}(x) = 1 - \frac{1}{2}(x-1)^2; \tag{4.43}$$

[1] One could of course choose a larger number of data points, however the rank of the covariance matrix of this set of data points will be m_0.

the true values x are transformed by the function

$$y_{tr} = x - 0.2\frac{x^2}{4} \tag{4.44}$$

to a variable y_{tr}, which is then assumed to be measured with a gaussian resolution function with a standard deviation $\sigma = 0.1$, resulting in the measured variable y. The assumed acceptance, the transformation function and the resolution function is shown in Figure 7 a, the response of the simulated detector to δ-function signals at $x = 0.5$, $x = 1.0$ and $x = 1.5$ is shown in Figure 7 b.

The assumed true function is

$$f(x) = \sum_{k=1}^{3} b_k \frac{g_k^2}{(x - x_k)^2 + g_k^2} \tag{4.45}$$

in the region $0 \leq x \leq 2$; the parameters used in the simulation are given in table 1.

k	b_k	x_k	g_k
1	1.0	0.4	2.0
2	10.0	0.8	0.2
3	5.0	1.5	0.2

Table 1. Parameters of true function.

A sample of 5000 random x-values is generated according to the function $f(x)$. A histogram of the sample is shown in Figure 8 a together with the function $f(x)$. After acceptance, transformation and smearing according to the assumptions made above 4475 y-values remain; a histogram of this sample, representing the result of the measurement, is shown together with the original function $f(x)$ in Figure 8 b.

The transformation restricts y_{tr} to $0 \leq y_{tr} \leq 1.8$, with the additional effect, that the two peaks become slightly narrower. The resolution function then broadens the peaks, filling up the valley between the peaks.

Now the unfolding method with regularization is applied to the data. The number of spline functions used in the discretization is $m = 22$. In table 2 several parameters of the unfolding method are shown, including the coefficients of the transformed solution. These values are also shown in Figure 9 a as bars. As is seen, about half of the coefficients are small and compatible with zero. The eigenvalues S_{jj} of the curvature matrix C, also given in table 2, increase by many orders of magnitude. A few of the orthogonal functions $p'_j(x)$ are shown in Figure 10. The amplitudes of the functions, each representing one standard deviation statistical error, increase with increasing index j. The regularization

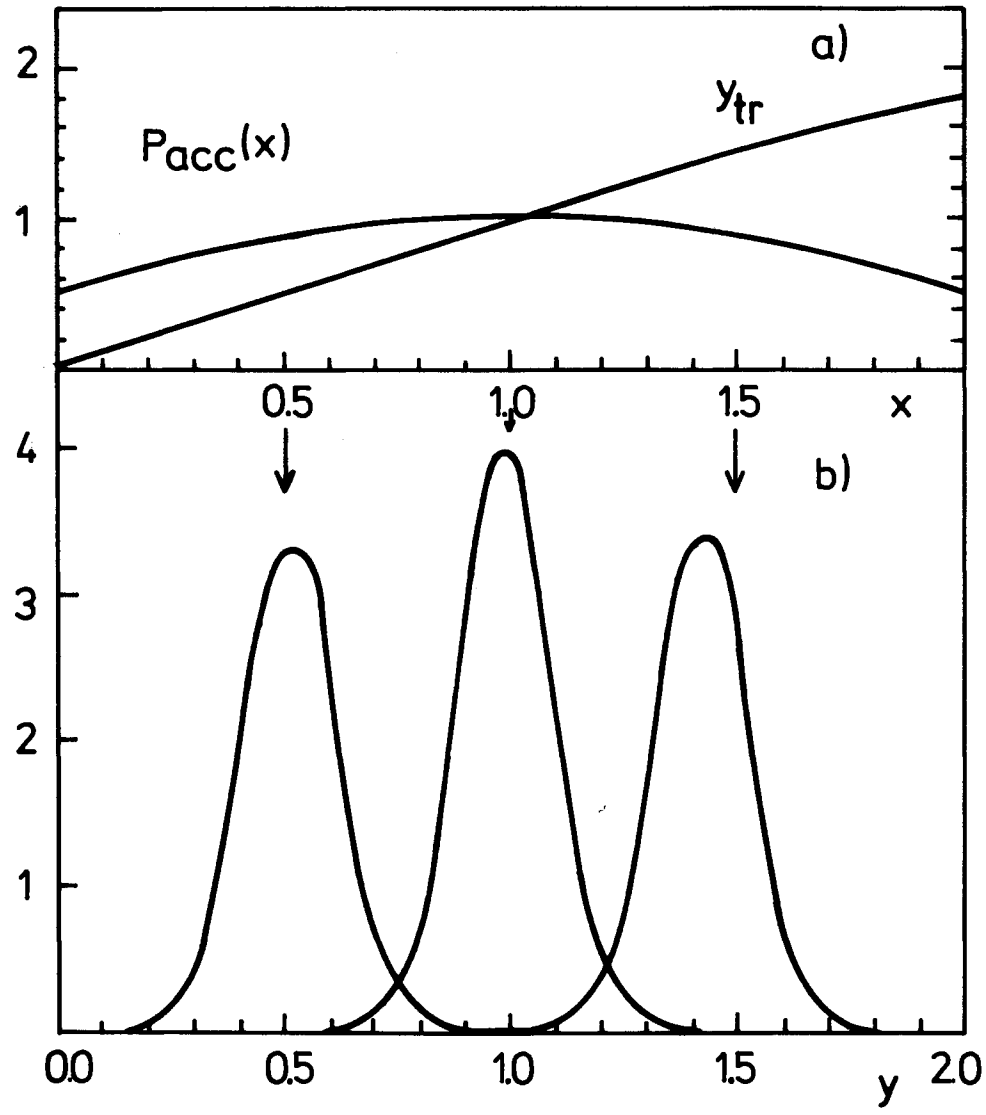

Figure 7. Acceptance, transformation and resolution, assumed in the simulation (a) and response to δ-functions at the values, indicate by arrows (b).

parameter τ is determined in this example for $m_0 = 12$, the result obtained from equation (4.41) is $\tau = 0.4287 \cdot 10^{-3}$. The coefficients of the regularized solution, obtained by multiplying the coefficients $\bar{a}'_j$ with the factors, as given by equation (4.40), and the factors themselves are given in the last two columns of table 2. The coefficients of the regularized solution are also shown in Figure 9 a. Figure 9 b shows the attenuation factor as a function of the index j.

Finally the set of coefficients is converted to data points, representing average values

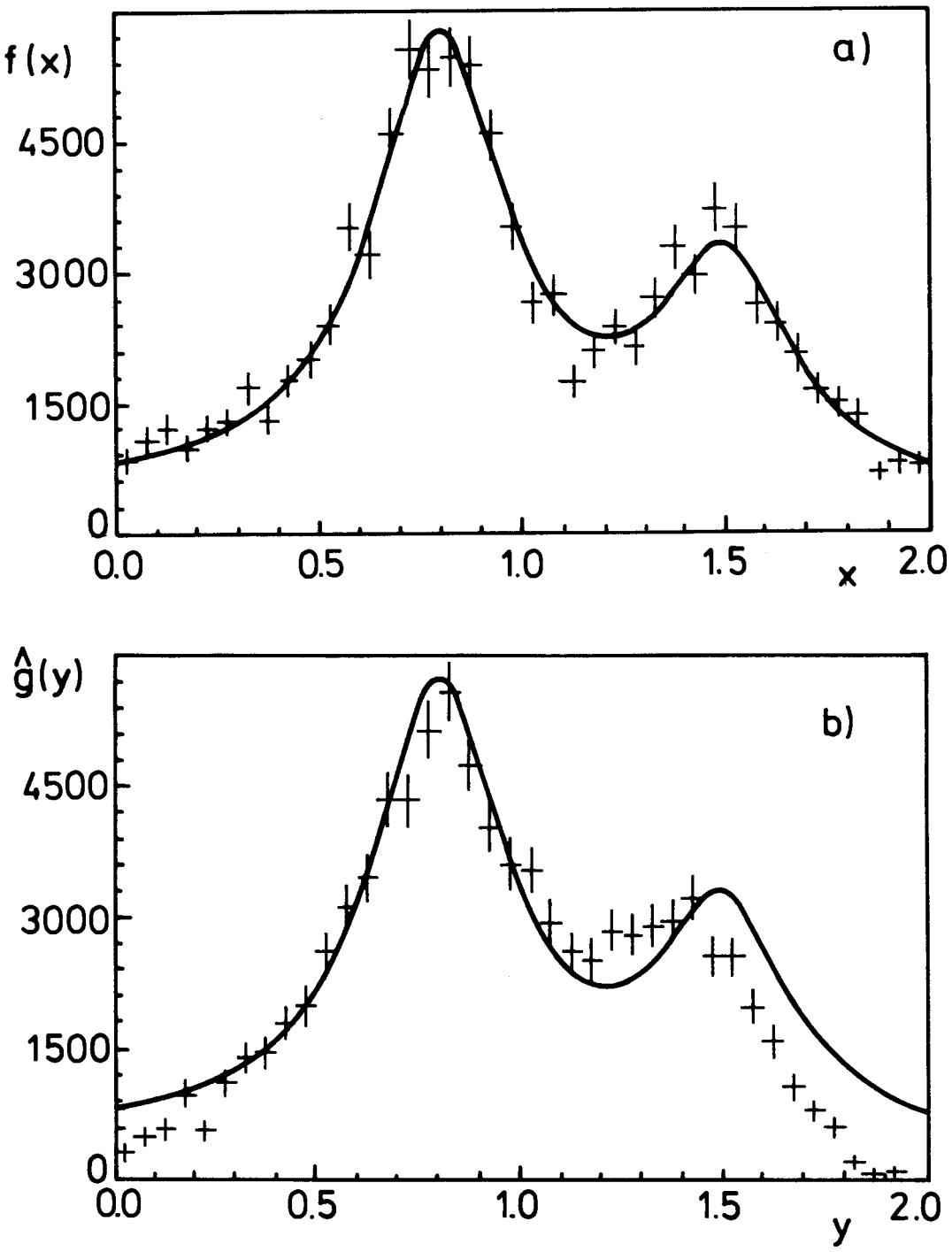

Figure 8. Histogram of the generated data (a) before and (b) after simulation of acceptance, transformation and resolution. The original function is shown as curve.

j	S_{jj}	$\bar{a}'_j$	$\hat{a}'_j$	factor
1	0.02	64.62	64.62	1.0000
2	0.09	1.70	1.70	1.0000
3	0.28	7.51	7.51	0.9999
4	0.79	9.86	9.85	0.9997
5	1.97	1.27	1.27	0.9992
6	4.86	11.23	11.21	0.9979
7	12.32	0.18	0.18	0.9947
8	29.61	3.72	3.67	0.9875
9	72.43	0.68	0.66	0.9699
10	183.90	0.04	0.04	0.9269
11	507.00	0.63	0.52	0.8214
12	1409.32	1.38	0.86	0.6233
13	$3.54 \cdot 10^3$	1.81	0.72	0.3974
14	$11.86 \cdot 10^3$	0.03	0.01	0.1644
15	$41.45 \cdot 10^3$	1.38	0.07	0.0533
16	$98.04 \cdot 10^3$	0.05	0.00	0.0232
17	$129.89 \cdot 10^3$	1.27	0.02	0.0176
18	$209.10 \cdot 10^3$	0.41	0.00	0.0110
19	$334.44 \cdot 10^3$	0.52	0.00	0.0069
20	$423.96 \cdot 10^3$	1.59	0.01	0.0055
21	$10.51 \cdot 10^6$	0.80	0.00	0.0002
22	$38.64 \cdot 10^6$	0.30	0.00	0.0001

Table 2. Parameters of the unfolding solution.

for the unfolded $f(x)$ in small x-regions. As explained above, the limits of these regions are determined by the positions of the extrema of the function $p'_{m_0+1}(x)$, in this case of the function $p'_{13}(x)$, which is shown in Figure 10 b, with the positions of the extrema indicated. As can be seen from Figure 10 b, due to the definition of the region limits the contribution of the function $p'_{13}(x)$ itself will be very small for the average values. The final result is shown in Figure 11 together with the original function $f(x)$. The unfolded data are within errors compatible with the function, in particular the valley between the peaks, hardly visible in the 'measured' data (Figure 8 b), is reproduced. Of course the statistical errors are much larger than in the histogram of the generated data in Figure 8 a. They illustrate the loss in statistical accuracy, that occurs due to a measurement with finite resolution.

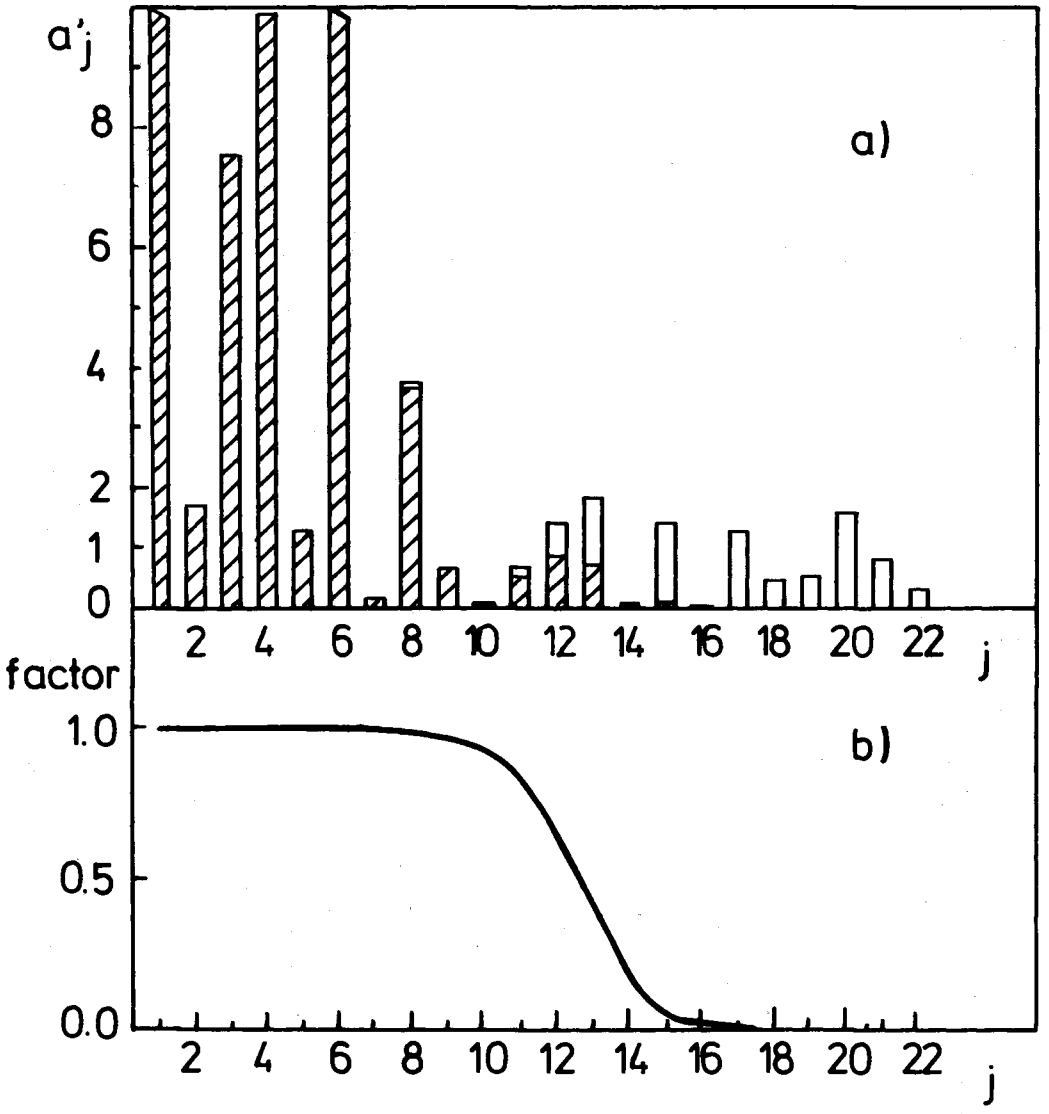

Figure 9. Plot of the coefficients of the normalized orthogonal functions of the solution (a); the dashed bars represent the values after the regularization. The attenuation factor is shown as a curve (b).

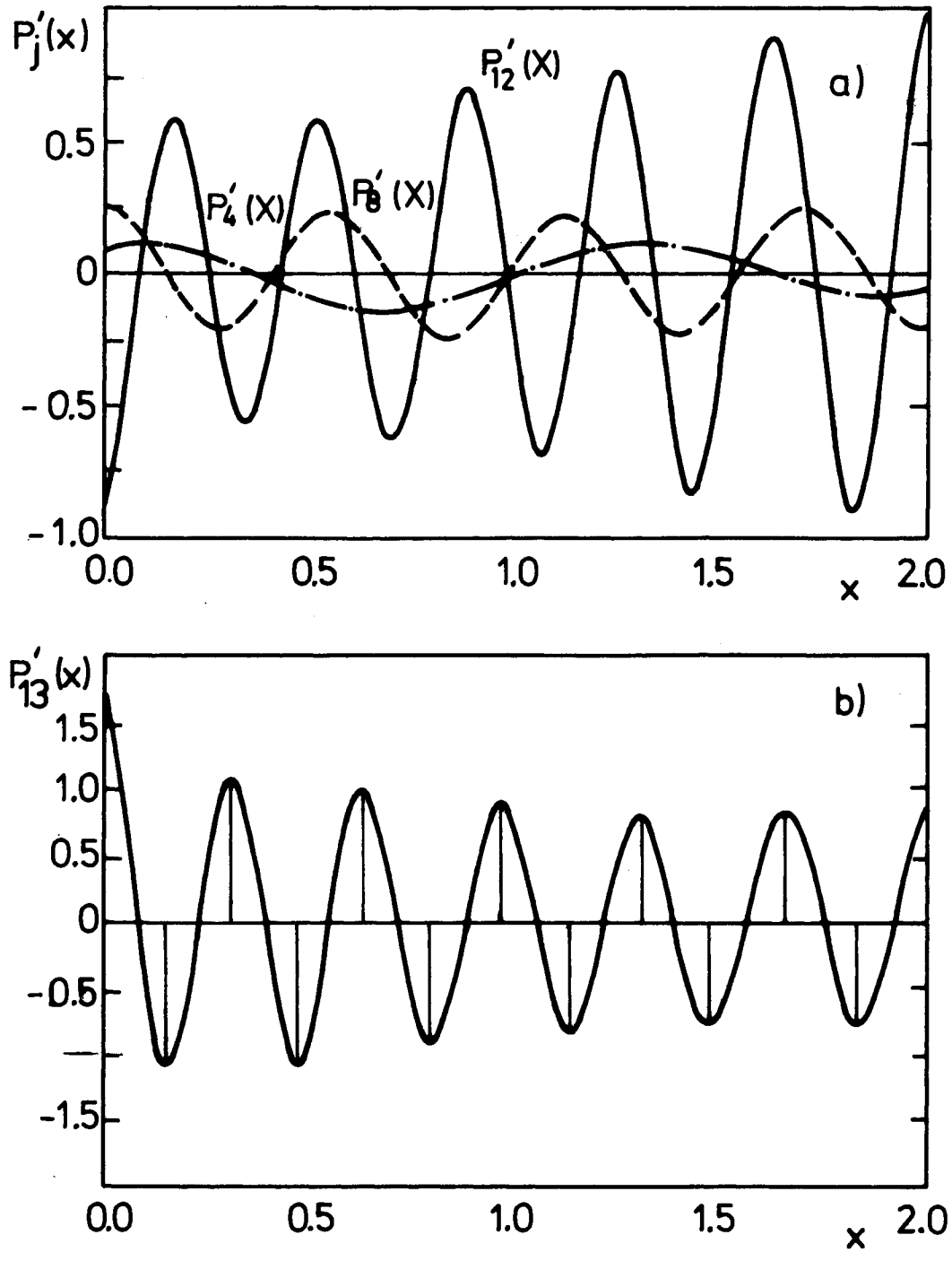

Figure 10. Examples for the normalized orthogonal functions, which are used to represent the solution.

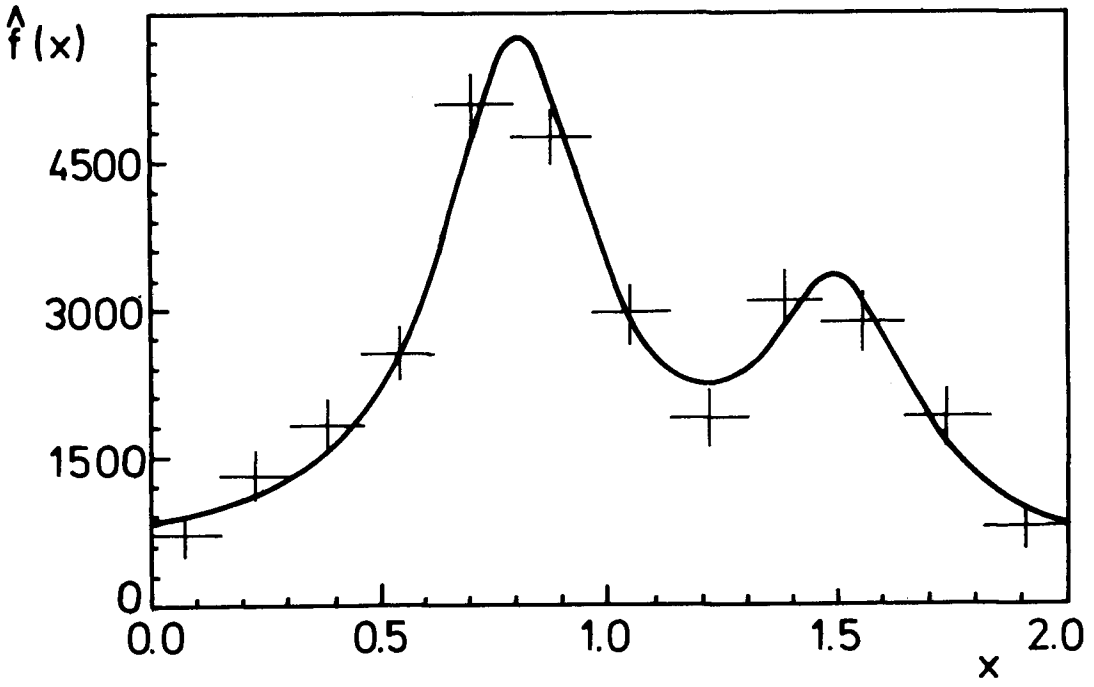

Figure 11. Result of unfolding with regularization, shown as data points together with the original function. The horizontal bar gives the range, over which the data points represents the average.

In conclusion, the regularization, which is the essential component of the unfolding method derived here, is shown to result in a smooth cut-off of insignificant higher order terms in the expansion of the unfolding solution in terms of orthogonal functions. The inclusion of these higher order terms would produce a spurious oscillatory component of the solution. Standard statistical tests are used for the determination of the regularization weight; they ensure that a possible bias introduced by the regularization is small compared to the statistical errors. The method also contains a prescription for the definition of a set $\{\hat{f}_k\}$ of average function values of the unfolded solution, which are only weakly correlated. It should be stressed, that unfolding cannot modify the statistical accuracy of an experiment, which is always reduced by the effect of limited resolution.

ACKNOWLEDGEMENT

The method discussed in chapter 4.4 of this paper has been applied to several high-energy physics experiments in an earlier and in the present version. For discussions and suggestions during the development I have to thank many collegues, in particular J. V. Allaby, L. Criegee, J. Dainton, U. Eckardt, R. Orr, K. H. Ranitzsch, I. Scillicorn and K. Winter. I want to thank L. Criegee and I. Scillicorn for a critical reading of the manuscript.

REFERENCES

1. J. H. Friedman, *Data analysis techniques for high energy particle physics*, in "Proceedings of the 1974 CERN School of Computing, Godoysund, Norway", CERN 74-23, 1974.

2. J. Carr, *Analysis of deep inelastic muon and electron scattering experiments*, in "Formulae and Methods in Experimental Data Evaluation with Special Emphasis on High Energy Physics", Vol. 2, European Physical Society, CERN, 1984.

3. I. P. Nedelkov, *Improper problems in computational physics*, Comp. Phys. Comm. **4** (1972), 157-164.

4. S. Christiansen, *Integral equations: An outline*, in "Formulae and Methods in Experimental Data Evaluation with Special Emphasis on High Energy Physics", Vol. 3, European Physical Society, CERN, 1984.

5. B. W. Rust and W. R. Burrus, "Mathematical Programming and the Numerical Solution of Linear Equations", American Elsevier Publishing Company, Inc., New York, 1972.

6. M. Jonker et al. (CHARM Collaboration), *Experimental study of differential cross sections $d\sigma/dy$ in neutral current neutrino and antineutrino interactions*, Physics Letters **102 B** (1981), 62-72.

7. Ch. Berger et al. (PLUTO Collaboration), *Measurement of the photon structure function $F_2^\gamma(x, Q^2)$*, Physics Letters **142 B** (1984), 111-118.

8. Ch. Berger et al. (PLUTO Collaboration), *Measurement of deep inelastic electron scattering off virtual photons*, Physics Letters **142 B** (1984), 119-124.

9. D. Drijard, *Design of experiments*, in "Proceedings of the 1978 CERN School of Computing, Jadwisin, Poland", CERN 78-13, 1978.

10. D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. Assoc. Comput. Mach. **9** (1962), 84-97.

11. A. N. Tikhonov and V. Ya. Arsenin, "Metody resheniya nekorrektnyh zadach", (Methods for Solution of ill-posed Problems), Nauka, Moscow, 1979.

12. J. Routti and J. V. Sandberg, *General purpose unfolding program LOUHI78 with linear and nonlinear regularization*, Comp. Phys. Comm. **21** (1980), 119-144.

13. S. W. Provencher, *A constrained regularization method for inverting data represented by a linear algebraic or integral equation*, Comp. Phys. Comm. **27** (1982), 213-227.

14. V. P. Zhigunov, *Improvement of resolution function as an inverse problem*, Nucl. Instrum. and Methods **216** (1983), 183-190.

15. M. Jonker et al. (CHARM Collaboration), *Experimental study of x-distributions in semileptonic neutral current neutrino interactions*, Physics Letters **128 B** (1983), 117-123.

16. W. T. Eadie, D. Drijard, F. E. James, M. Roos and B. Sadoulet, "Statistical Methods in Experimental Physics", North-Holland Publishing Company, Amsterdam, London, 1971.

17. V. Blobel, *Least squares methods*, in "Formulae and Methods in Experimental Data Evaluation with Special Emphasis on High Energy Physics", Vol. 3, European Physical Society, CERN, 1984.

18. J. Stoer and R. Bulirsch, "Introduction to Numerical Analysis", Springer-Verlag, New York, Heidelberg, Berlin, 1980.

19. H. Wind, *Interpolation and function representation*, in "Formulae and Methods in Experimental Data Evaluation with Special Emphasis on High Energy Physics", Vol. 3, European Physical Society, CERN, 1984.

20. C. de Boor, "A Practical Guide to Splines", Springer-Verlag, New York, Heidelberg, Berlin, 1978.

21. R. Courant und D. Hilbert, "Methoden der Mathematischen Physik I", Springer-Verlag, Berlin, 1924.

22. C. Lanczos, "Applied Analysis", Prentice-Hall, Inc., Englewood Cliffs, N.J., 1956.

23. E. L. Kosarev and E. Pantos, *Optimal smoothing of noisy data by fast fourier transform*, Sci. Instrum. **16** (1983), 537-543.