Published by Institute of Physics Publishing for SISSA

Received: *October 5, 2005*
Revised: *June 16, 2006*
Accepted: *July 4, 2006*
Published: *August 24, 2006*

Smoothing of one and two-dimensional histograms with a diffusion algorithm

Peter Bock

Physikalisches Institut der Universität
Philosophenweg 12, D69120 Heidelberg, Germany
E-mail: peter.bock@physi.uni-heidelberg.de

Abstract: Monte Carlo histograms with statistical fluctuations and without proper mathematical modelling are often used to analyse experimental data. A procedure to smooth one and two-dimensional histograms is described which uses a diffusion-like information exchange between neighboring bins to average over statistical fluctuations. The procedure is steered by two parameters which control the fraction of the total histogram content to be used in the local averaging process.

Keywords: Statistical Methods, Higgs Physics.

JHEP08(2006)056

Contents

JHEP08(2006)056

1. Introduction

Many analyses of complex experimental data use Monte Carlo predictions for observables to adjust theory parameters to observed distributions. Often no mathematical models for histogram shapes exist, examples being the treatment of instrumental responses or plots of likelihood and neural network variables. Sometimes one prefers a smooth behavior of theoretical distributions, if statistical errors have to be computed from a histogram for fits, or interpolations between distributions have to be performed. In particle search experiments, the statistical analysis is sometimes based on the likelihood ratio between a hypothetical signal distribution and an underlying flat background [1]. A background histogram may have empty bins, and if one of these contains an observed event, the computed likelihood ratio becomes infinite.

In this situation, the true distribution has to be estimated from a histogram by averaging out the statistical fluctuations, without detailed *a priori* knowledge of its correct shape.

Any acceptable smoothing algorithm should fulfill the following criteria:

- Statistically insignificant fluctuations have to be removed.
- Significant structures in high intensity bins have to be be kept with distortions as small as possible.

- Many physical observables like angles or neural network variables have lower and upper bounds and no events lie outside the histogrammed region. The total rate is often predicted by the Monte Carlo computation producing the histogram. The overall normalization of the histogram should then be fixed.

There are many publications on smoothing algorithms. Procedures popular in high energy physics are spline fits and multiquadric smoothing [2, 3]. In the latter case, centers of a distribution are searched for where the second differential becomes statistically significant. The distribution is approximated by basis functions depending on the distance of a point from its relevant center and a curvature. The algorithm is steered by two parameters controlling the setting of the statistical sensitivity and the curvatures. Spline fits need two steering parameters, too, which are the number of knots and the degree of spline functions.

An alternative method, presented in ref. [4], is the description of the data by Gaussian kernels. The algorithm depends on one parameter only. It uses all data events individually, generates an interpolating function and is thus free from histogram binning effects.

In the search experiments mentioned above, the problem of fluctuations in low intensity background distributions may be a crucial one. A critical comparison of some existing smoothing algorithms has shown that they gave unsatisfactory results, with the exception of the Gaussian kernel method [5].

In this work, an alternative procedure fulfilling the above criteria is described: a diffusion algorithm with a bin dependent diffusion constant adjusted to the local intensities. The method is closely related to the method of Gaussian kernels and the numerical results are quite similar. The program uses binned histograms as input information and is thus less ambitious than the Gaussian kernel method in a technical sense. It is, on the other side, more general than the implementation described in ref. [4], because it contains a second steering parameter to control the statistical fluctuations after smoothing. It is the aim of this paper to study the impact of the additional parameter on the systematic errors of the smoothing algorithm.

The method was introduced for complex Higgs search analyses [7], where experimental results for several accelerator energies and decay channels had to be combined. In addition, Monte Carlo predictions had to be made for a set of hypothetical Higgs masses. In total, many hundreds of mass and likelihood distributions had to be handled. It was the aim to find a simple equivalent for the Gaussian kernel method, which converts a histogram directly into a smoothened histogram and avoids the need for many function routines, thus simplifying code management. Instead, all information on distributions is stored in a data base.

The code has been developed for one- and two-dimensional distributions. The construction of the algorithm is by far not unique. Three different variants were investigated in detail and their performance and the systematic errors were compared.

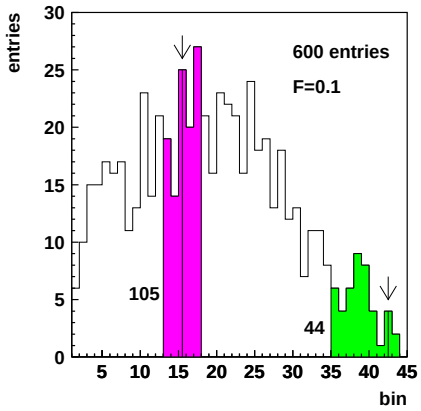

Figure 1: Initial averaging. The arrows mark two bin positions where v_i is evaluated.

2. Smoothing of one-dimensional histograms

2.1 Control of statistical fluctuations

It is assumed that the errors per bin are proportional to $\sqrt{r_i}$, where r_i is the true intensity. Any smoothing procedure requires an averaging over a number of bins, $n_{eff,i}$. In an initial pre-smoothing step, a constant weighting is introduced. The histogram smoothing will produce a modified set of rates r_i^*, and the relative errors become

$$\frac{\delta r_i^*}{r_i^*} \sim \frac{1}{\sqrt{n_{eff,i} \cdot r_i^*}} \tag{2.1}$$

As a parameter which controls the residual statistical fluctuations at the histogram maximum after smoothing, a predefined number of bins, $n_{eff,max}$, could have been introduced. Equivalently, the ratio

$$F = \frac{n_{eff,max} \cdot r^*_{\max}}{\sum_{k=1}^{n_b} r_k} \tag{2.2}$$

is used throughout this paper. The sum in the denominator extends over all n_b bins of the histogram, so that F is defined as the fraction of the total histogram content to be used in the averaging process around the maximum.

Additionally, one has to specify the statistical fluctuations at arbitrary histogram bins. The ansatz taken is a power law as the function of the local rate:

$$\frac{\delta r_i^*}{r_i^*} = \frac{\delta r^*_{\max}}{r^*_{\max}} \cdot (\frac{r^*_{\max}}{r_i^*})^{\kappa} \tag{2.3}$$

Both F and κ have to be specified by the user. They have to be chosen in such a way that those peaks considered as significant are not averaged away by the pre-smoothing. The exponent κ lies between the boundaries $\kappa = 1/2$ and $\kappa = 0$. In the first case, the

error (2.1) follows the original square root law and the number of bins $n_{eff,i}$ is the same everywhere in the histogram. The smoothing corresponds to an overlay of some histograms with a coarser binning, shifted against each other. The other limit realizes the extreme case that the relative error is the same in all histogram bins after smoothing, which makes $n_{eff,i}$ strongly bin dependent. The systematic errors are smallest in the first and largest in the second case.

A compromise between these extremes, suitable as a default, is $\kappa = 1/4$. It should be noted that the implementation of the Gaussian kernel method of ref. [4] always uses a kernel width proportional to $1/\sqrt{r_i^*}$ as proposed in reference [6], which is equivalent to this κ value. Throughout this paper, the algorithms based on $\kappa = 1/4$ and $\kappa = 0$ are referred to as weak and strong smoothing, respectively.

Equation (2.3) gives, together with the error (2.1) and the definition (2.2):

$$\frac{n_{eff,i} \cdot r_i^*}{n_{eff,max} \cdot r_{\max}^*} = (\frac{r_i^*}{r_{\max}^*})^{2\kappa} \tag{2.4}$$

$$n_{eff,i} \cdot r_i^* = F \cdot (\sum_{k=1}^{n_b} r_k) \cdot (\frac{r_i^*}{r_{\max}^*})^{2\kappa} \tag{2.5}$$

The real bin contents are unknown. An estimate v_i for it is extracted from the histogram with the definition

$$v_i = \frac{\sum_{k=\max(i-j,1)}^{\min(i+j,n_b)} h_k}{n_{eff,i}^*} \tag{2.6}$$

with the number of bins

$$n_{eff,i}^* = \min(i+j, n_b) - \max(i-j, +1) + 1\,, \tag{2.7}$$

where j is an integer number.

To get compatibility of $n_{eff,i}^*$ with eq. (2.5), the free parameter j is set to the lowest value which fulfills the inequality

$$\sum_{k=\max(i-j,1)}^{\min(i+j,n_b)} h_k \geq (F \cdot \frac{n_{eff,i}^*}{2j+1}) \cdot (\sum_{1}^{n_b} h_k) \cdot (\frac{v_i}{v_{\max}})^{2\kappa} \tag{2.8}$$

with $v_i > 0$. This condition, together with eq. (2.6), always gives a unique solution. Normally, $n_{eff,i}^* = 2j+1$ bins contribute to the sum on the left hand side. Close to the histogram boundaries, the counting can be truncated at one side and the factor F is reduced automatically. A larger statistical rest fluctuation is tolerated near the boundaries with the consequence that v_i is a better approximation for the true rate.

The maximum rate $v_{\max}$ does not depend on κ and has to be evaluated first. The construction is illustrated in figure 1 for $\kappa = 0$. The spectrum v_i cannot be taken as the final result. Its disadvantage is the use of an equal weighting for all bins. Also the number of events is not conserved. A better performance can be reached with a Gaussian weight profile.

2.2 The algorithm

Gaussian smearing can be approximated by a stepwise information exchange between neighboring bins of the histogram. If h_{i-1}, h_i and h_{i+1} are the contents of 3 adjacent bins, the intensity in bin i will be modified according to

$$h_i^* = h_i \cdot (1 - f_{i,i-1} - f_{i,i+1}) + h_{i-1} \cdot f_{i,i-1} + h_{i+1} \cdot f_{i,i+1} \ . \tag{2.9}$$

The $f_{i,i\pm1}$ are exchange coefficients to be defined later.

At the histogram boundaries one sets, for the non-existing bins 0 and $n_b + 1$,

$$f_{0,1} = 0 \qquad f_{n_b,n_b+1} = 0 \ . \tag{2.10}$$

A natural upper limit for the $f_{i,k}$ coefficients is obtained for the case that the content of a bin is equally distributed over 3 bins:

$$\max(f_{i,k}) = f_{\max} = \frac{1}{3} \ . \tag{2.11}$$

The mixing process (2.9) is done in parallel for all bins and it is repeated until a certain number of iterations N is reached. The factors $f_{\rm ik}$ are adjusted to the local rates in such a way that the number of steps N is universal for a histogram.

For one selected histogram bin i and fixed f, the distribution obtained after N iterations quickly approaches a Gaussian distribution with the variance

$$\sigma_i^2 = 2 \cdot f \cdot N \, , \tag{2.12}$$

where σ_i is measured in number of histogram bins. To get consistency with the parameterization of the preceding subsection, the variance (2.12) is made equal to the variance of the $n_{eff,i}$ bins for $n_{eff,i} \gg 1$:

$$\sigma_i = \frac{n_{eff,i}}{\sqrt{12}} \ . \tag{2.13}$$

The largest value of f is assigned to the bin with the lowest intensity. Equations (2.13), (2.12) and (2.5) are then sufficient to compute the number of steps, with r_i^* replaced by v_i:

$$N = \frac{1}{24 \cdot f_{\max}} \cdot F^2 \cdot (\sum_{k=1}^{n_b} h_k)^2 \cdot (\frac{v_{\min}}{v_{\max}})^{4\kappa} \cdot \frac{1}{v_{\min}^2} \tag{2.14}$$

For the exchange coefficients one gets the proportionality

$$f \sim {v_i}^{4\kappa-2} \tag{2.15}$$

In the application, the f factors have to be assigned to pairs of adjacent bins, which leads to the ansatz

$$f_{i,i\pm1} = f_{\max} \cdot (\frac{2 \cdot v_{\min}}{v_i + v_{i\pm1}})^{2-4\kappa} \ . \tag{2.16}$$

The request for a minimal fraction F avoids divergencies in the last equation.

The iteration formula (2.9), the exchange coefficients (2.16) and the number of steps (2.14) define the algorithm for one dimensional smoothing. It depends on the two parameters F and κ and contains the v_i, which are estimates for the unknown true rates r_i and were defined with the initial averaging process (2.8), (2.6).

It has been assumed implicitly that this auxiliary spectrum is generated with the same parameters F and κ as used in the main analysis. In the following, this is called the self consistent approach. This is not mandatory. As will be shown, the systematic errors can be reduced a bit, if the v_i are computed with a smaller κ value, at the price of somewhat larger statistical rest fluctuations at low rates. Examples were studied with $\kappa = 0$ for the pre-smoothing and $\kappa = 1/4$ for the main analysis with the same parameter F; this parameter combination is called the 'mixed approach'.

2.3 Performance

The differences in performance between the algorithm described here and available standard routines [2] are illustrated in figure 2. The underlying model for the Monte Carlo data set consists of two Gaussians with a peak rate of 100 events per bin and a width of 4 bins, superimposed on a constant background of 5 events per bin. The upper right picture gives the result of the diffusion algorithm with $F = 0.15$. Because the input histogram contains 2000 entries, this corresponds to a smearing over 3 bins and a residual statistical fluctuation of $(1/\sqrt{300})$=5.8% at the maximum. The input data have a statistical fluctuation of more than one standard deviation at the central peak; a part of this structure is left over after smoothing. The background level is well reproduced with some remaining structures, which are clearly correlated to statistical fluctuations in the input data. With the default parameter $\kappa = 0.25$, statistical oscillations with amplitudes of the order of $\sqrt[4]{100/5}\cdot 5.8\% \approx$ 12% and correlation lengths of 6 bins are expected, in agreement with the figure. A slight broadening of the Gaussian peaks is visible. The pedestal can be made more flat and the structure at the top can be removed with $F = 0.20$, at the price of a slightly increased peak broadening. The lower left curve shows the result from the multiquadric approach. The fluctuations in the pedestal region are larger and its correlations to the input data are not always obvious. The shapes of the main peaks are somewhat non-parabolic in logarithmic representation. An alternative method of smoothing from ref. [2] is the so called 353QH algorithm. As shown in the lower right picture, the two peaks are well reproduced, but the method creates, on the other hand, sudden jumps in the low intensity region. The spline smoothing, not shown at all, needs special tuning to reproduce the peak width and introduces then unacceptable oscillations in the pedestal region, too. Instabilities of this type are the reason for the problems in the application mentioned in the introduction.

The comparison shows the advantage of the diffusion algorithm: The residual statistical fluctuations can be controlled directly with the parameters F and κ over a dynamic range of more than one order of magnitude. A disadvantage is an unwanted, but unavoidable feature that major structures are widened. The systematic errors are proportional to F^2. They may become significant, if the algorithm is applied to a histogram with low statistics and a large dynamic range; especially if the intensity goes to zero. In this situation there is a conflict between the requirements of smoothness and the correct reproduction of the

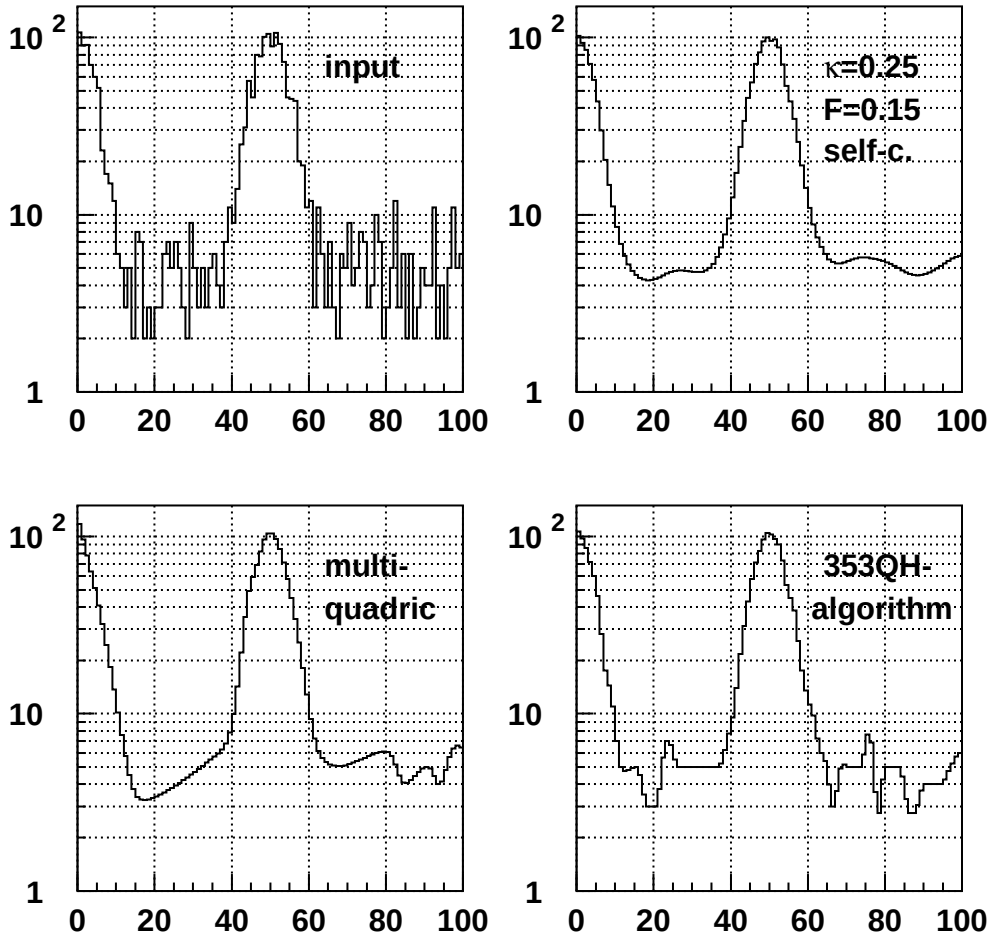

Figure 2: it Smoothing of a data set with different algorithms. Upper left: input data, upper right: this work, lower part: multiquadric procedure (left) and 353QH algorithm (right) of ref. [2] with default parameters.

underlying distribution: A smooth result requires a large F value, to be estimated from the allowed residual statistical fluctuations; a limit on systematic errors will set, however, upper bounds on F, as explained in the next section.

2.4 Systematic errors

The diffusion algorithm does not converge in a statistical sense: If the smoothing is repeated many times with fluctuating input histograms, the averaged smoothened histogram is not identical with the real distribution. Only the expectation value of the sum of histogram

entries is correctly reproduced, because the sum of entries is kept by construction. The systematic errors depend, apart from statistical second order effects from pre-smoothing, on the parameters F and κ only and not on the number of histogram entries.

Peak broadening is most easily estimated for strong smoothing. Let n_p be the intensity of a Gaussian peak, superimposed on a constant background with b events per bin, and σ_p the peak width. The rate at half maximum is $v_p = b + n_p/(2\sqrt{2\pi} \cdot \sigma_p)$. The variance of peak broadening is, according to eqs. (2.12), (2.14) and (2.16),

$$2 \cdot f \cdot N \approx 2 \cdot F^2 \cdot \frac{n^2 \cdot \sigma_p^2}{(n_p + 2\sqrt{2\pi} \cdot b \cdot \sigma_p)^2} \ . \tag{2.17}$$

If the background is negligible, the broadening $\sqrt{2fN}$ is proportional to the width σ_p, irrespective of the binning. The effect will be small as long as $F \ll n_p/n$. The constraint is a bit less severe for weak smoothing, where an additional factor $\sqrt{v_p}/\sqrt{v_{\max}} < 1$ appears in the broadening. The distortions are at the % level, if $F < 0.1 \ldots 0.2 \cdot n_p/n$. If the rate in the peak region is dominated by background, the broadening becomes independent of the width; little narrow structures will always be removed.

There are two additional imperfections of the algorithm: Because the width of smearing σ_i is largest at lowest intensities, the smoothing creates peak tails, in addition to the mean broadening already discussed. Furthermore, condition (2.10) introduces 'event reflections' at both boundaries of the histogram. The algorithm has therefore the tendency to move events from the interior of the histogram to a boundary or vice versa, depending on the sign of the end slope of the distribution. As a result, the reconstructed end slopes are too small. In case of very small intensities this can mimic pedestals, which are, however, local accumulations of events with the compensating deficits spread over a wider histogram region.

To get upper limits for systematic errors, these effects were studied in detail for a Gaussian and a linear distribution. The numerical examples are constructed as worst case scenarios for the diffusion algorithm: The intensities per bin were less than 15 and the largest fractions F were 0.2. The true intensities were set to zero at the histogram boundaries, where the tail and reflection effects accumulate. Especially, the dependence of the errors on the κ parameter was investigated by changing κ from its default value to the lower limit $\kappa = 0$, which gives the largest systematic errors.

Normal distribution As an example, figure 3 gives the result obtained with an input data set generated by applying statistical fluctuations to a Gaussian function with a width of 10 histogram bins (upper left picture).

A dip due to a large statistical fluctuation at the maximum of the distribution survives after smoothing. The differences between the two weak smoothing algorithms are marginal for the special example and only the mixed approach results are shown in figure 3. To study the systematic errors quantitatively, 10000 Monte Carlo distributions, each consisting of 1000 events, were generated from the Gaussian function in figure 3. Figure 4 shows the systematic differences between the averaged smoothened histograms h_{mean} and the real distribution h_{true}. The differences are normalized to the true maximum $\max(h_{\text{true}})$. The

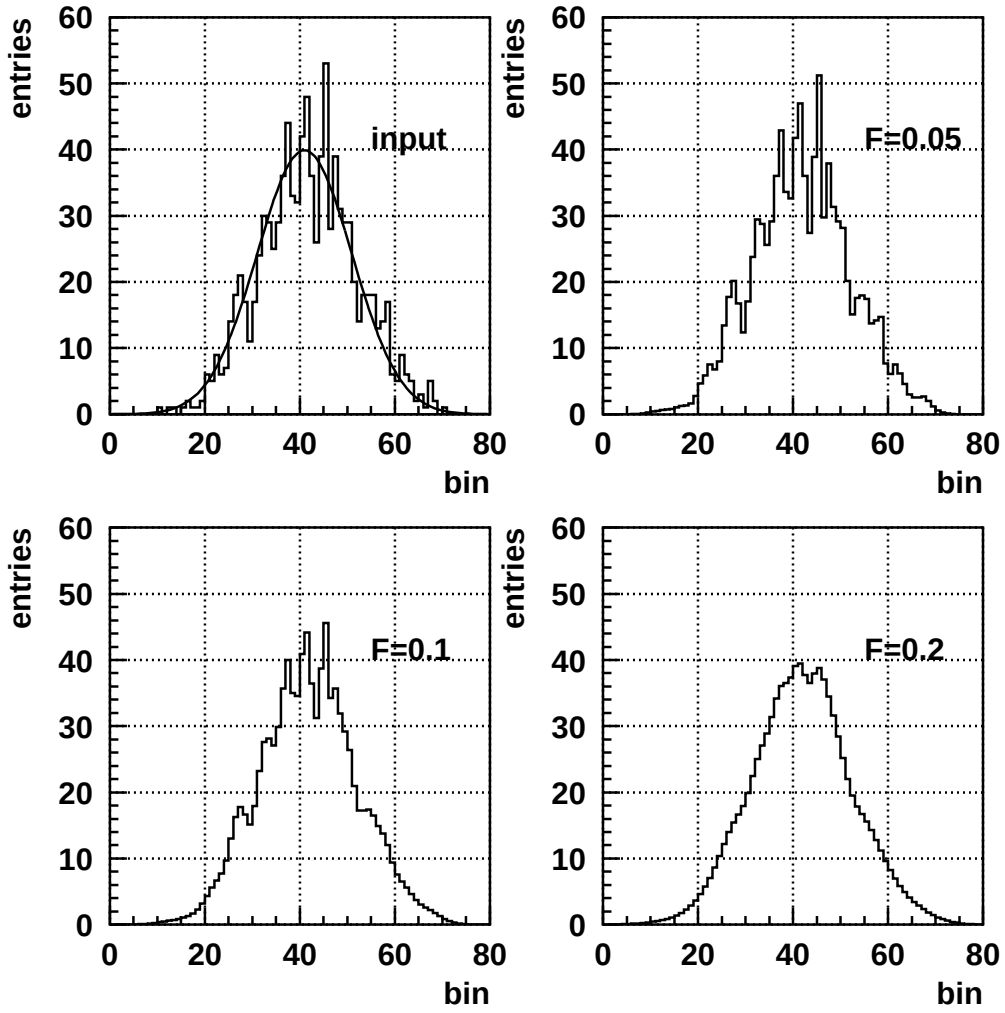

Figure 3: Smoothing of a Gaussian distribution with $\kappa = 0.25$ and different values of the parameter F. The total number of entries in the histogram is 1000, the bin-to-bin fluctuations follow the polynomial distribution.

tails of the true distribution are overlayed for comparison. For large fractions F, significant pedestals appear at the histogram boundaries, especially for strong smoothing.

Figure 5 gives information about the cumulated distributions. Again the averaged reconstructed histograms are compared with the original Gaussian function. A formal pull $p = p(j)$ is defined for every bin j by relating the summed histogram entries to a Gaussian integral:

$$\frac{\sum_{i=1}^{j} h_{mean,i}}{\sum_{i=1}^{n_b} h_{mean,i}} = \int_{-\infty}^{p(j)} \frac{1}{\sqrt{2\pi}} \exp(-\xi^2/2) d\xi \ . \tag{2.18}$$

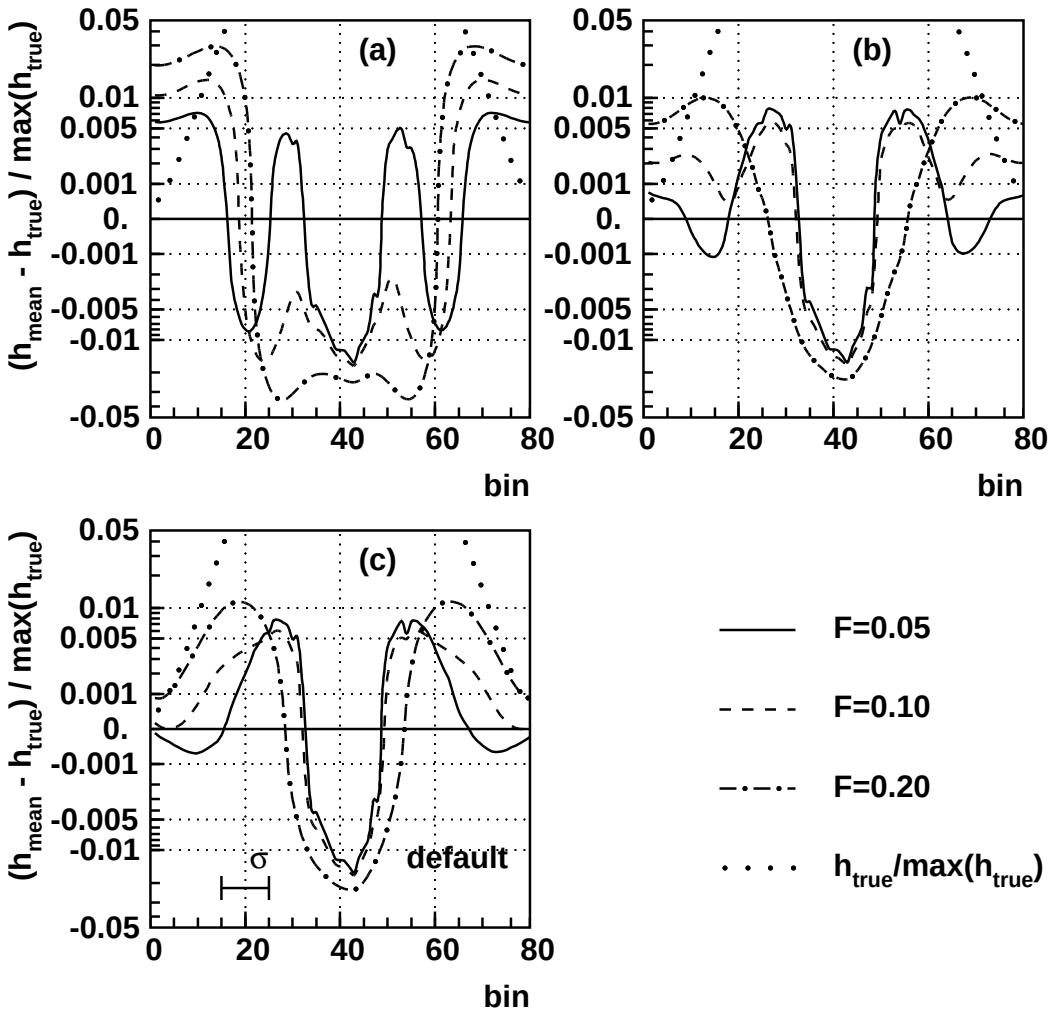

Figure 4: Difference of the mean smoothened histogram h_{mean} from an original Gaussian function. The number of histogram bins per standard deviation is $\mu = 10$. (a): strong smoothing, (b) weak smoothing, self consistent approach, (c) weak smoothing, mixed approach. The dots mark the original function. All function values z_i were transformed to a non-linear scale with $\pm \ln(1 \pm z_i/0.001)$.

The same procedure is applied to the original Gaussian histogram. Instead of the summed histogram entries, the equivalent pulls, computed with eq. (2.18), are used to prepare figure 5, with the limiting bin j as running parameter. The pull ratio between the averaged reconstructed and the true distributions measures the relative local widening due to the smoothing. This presentation of results depends only weakly on the binning and the number of histogram entries.

The central parts of the curves are flat, the onsets of major slopes being F dependent. It

JHEP08(2006)056

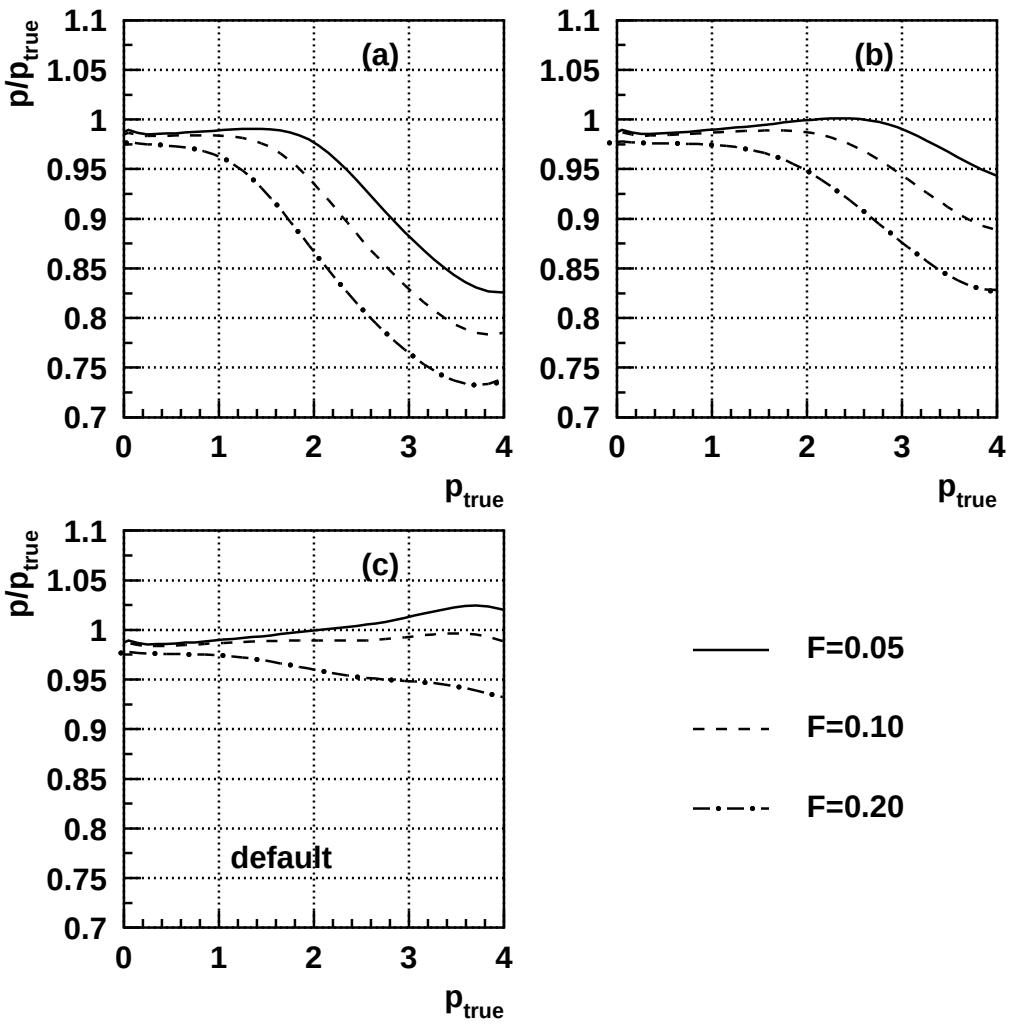

Figure 5: Ratios between mean reconstructed and true pulls of a Gaussian distribution as a function of the true pull. (a): strong smoothing, (b): weak smoothing, self consistent approach, (c): weak smoothing, mixed approach.

is evident that weak smoothing has a better tail behavior. Comparison of the self consistent approach with the mixed approach shows, that the former ansatz produces larger tails. The mixed approach turns out to be the best compromise between the requirements of smooth results and small systematic errors and is therefore taken as the default.

Linear function A Monte Carlo data set was generated by applying statistical fluctuations to a linear distribution which vanishes at $x = 0$. The expectation value for the number of entries in the uppermost bin is 10. The result obtained with weak smoothing is presented in figure 6. With the parameter $F = 0.2$ the number of events contributing to the

JHEP08(2006)056

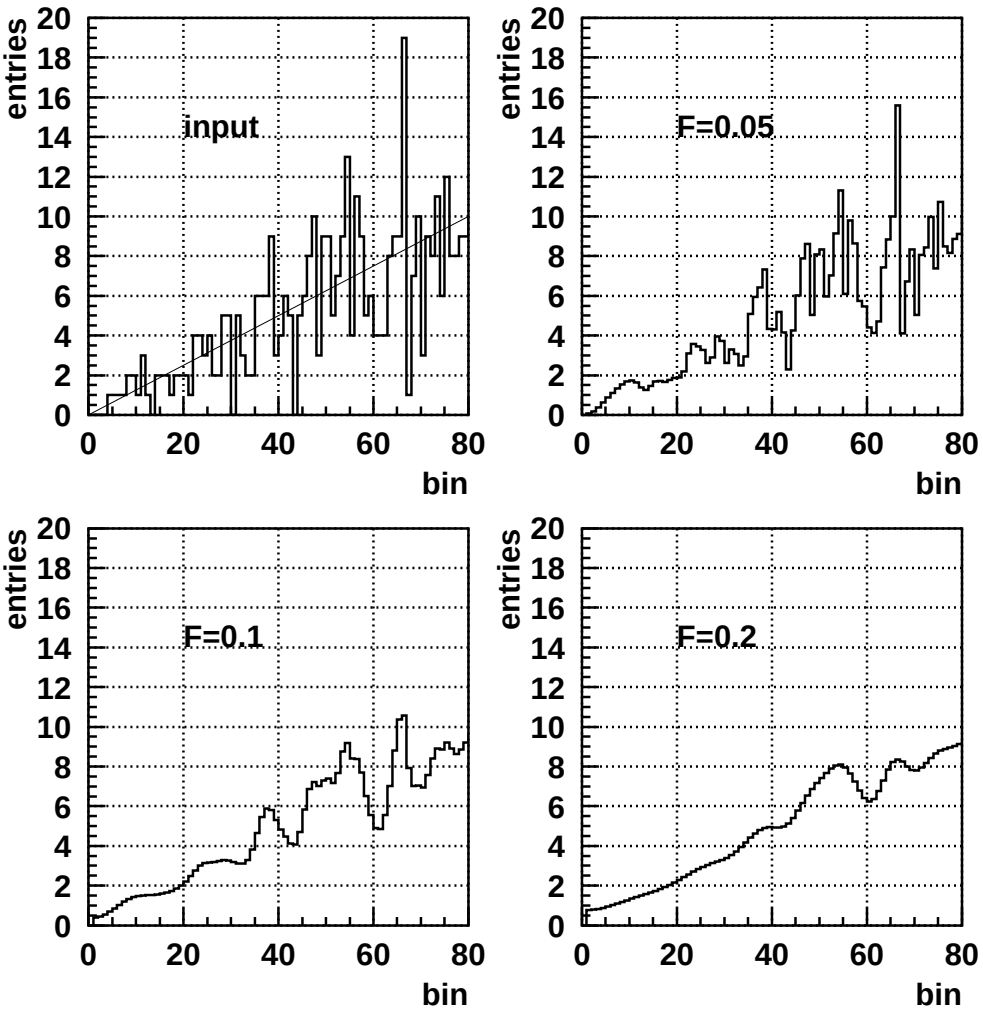

Figure 6: Smoothing of a Monte Carlo histogram, generated from a linear function, with $\kappa = 0.25$ and different values of the parameter F. The total number of entries in the histogram is 400.

smoothened histogram in the uppermost bin is 80 and the typical amplitude of statistical oscillations is therefore $\approx 11\%$.

Figure 7 shows the differences between the mean reconstructed distributions and the true one, obtained from 50000 Monte Carlo data sets. The differences are normalized to the true maximum value at the upper boundary.

For any value of F the mixed approach version of weak smoothing gives the lowest systematic errors, as for the Gaussian case.

Delta function Histograms with only one filled bin contradict the assumption of smooth variation of the true rate r_i with i, which has been implicitly made at the beginning. If a

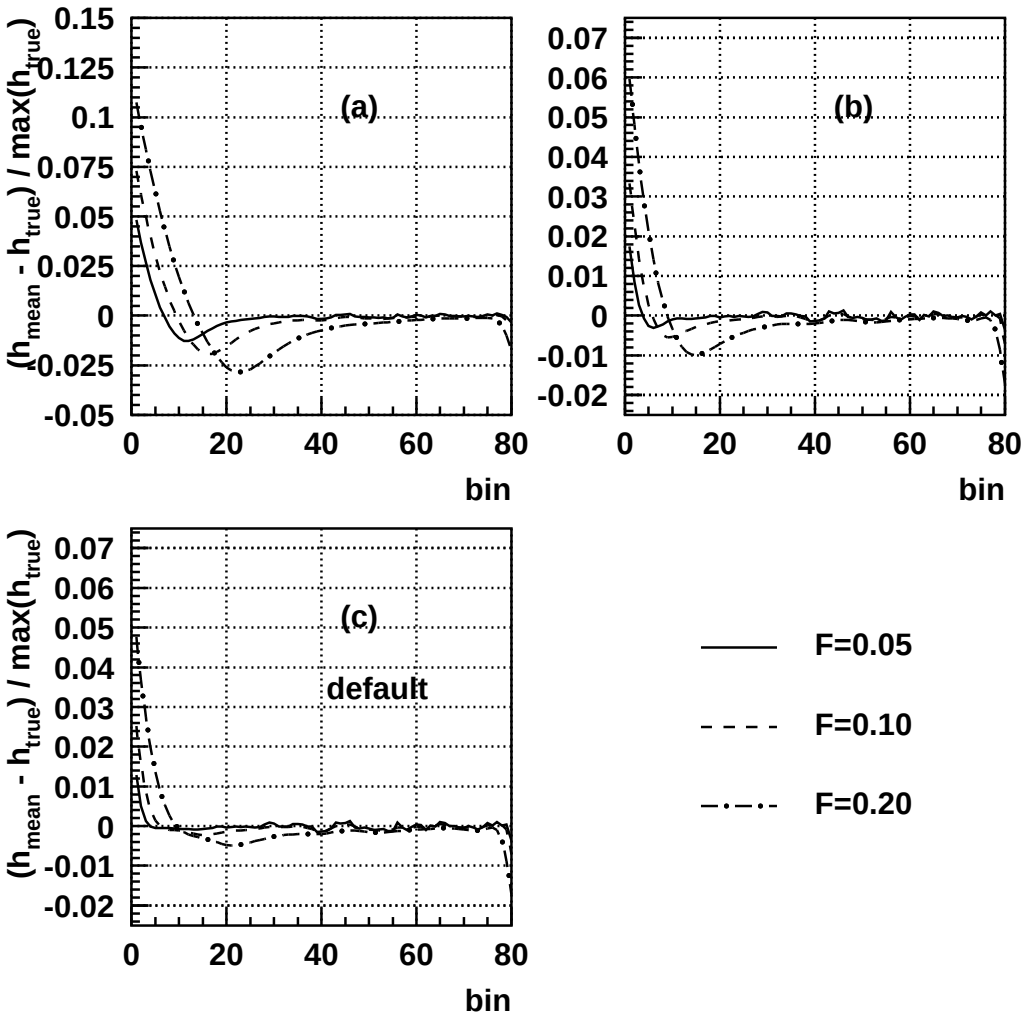

Figure 7: Average reconstructions of a straight line as a function of the the parameter F. The true distribution is taken from figure 6. (a): strong smoothing, (b): weak smoothing, self consistent approach, (c) weak smoothing, mixed approach.

one bin spike appears, the algorithm produces a tail around the filled bin, which has to be considered as a systematic error.

Again, the error is largest for strong smoothing. The diffusion causes an intensity loss in the bin originally filled, which appears as the total tail intensity. If backward diffusion is neglected, the relative intensity loss is given by $1-(1-2f)^N$ with the diffusion coefficient from eq. (2.16):

$$f = \frac{4}{3} \cdot \frac{v_{\min}^2}{\max(v_i + v_{i+1})^2} \ . \tag{2.19}$$

JHEP08(2006)056

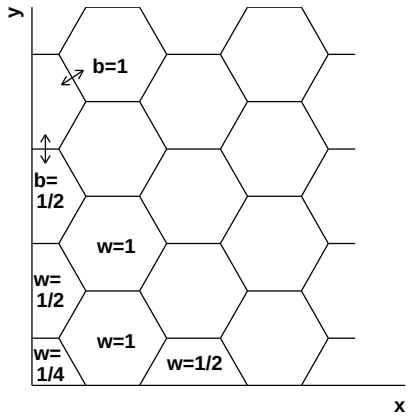

Figure 8: Hexagonal binning of 2-dimensional histograms. The parameters w and b measure the cell sizes and the lengths of common boundary lines of two cells.

For sufficiently small F, no smoothing is done, because eq. (2.14) gives $N = 0$. If $F > 1/n_b$, N may become large. For the delta function one gets, with eq. (2.6)

$$\max(v_i + v_{i+1}) = \sum_k h_k \cdot (1 + \frac{1}{3}) \,, \tag{2.20}$$

and with the number of steps from eq. (2.14), the relative tail intensity becomes

$$1 - (1 - 2f)^N \approx 2Nf = \frac{3}{16} F^2 \,, \tag{2.21}$$

which is less than 1% up to $F = 0.2$.

3. Smoothing of two-dimensional histograms

3.1 Hexagonal binning

For geometrical reasons two-dimensional bins have to be triangular, rectangular or hexagonal, if all bins are requested to have the same shape and size. Diffusion of a delta-function leads to a two-dimensional Gaussian. Any binning breaks the rotational symmetry after diffusion, the smallest periodicity angle in Euclidian geometry being 60 degrees for regular hexagons. Triangular and rectangular bins have the additional disadvantage of having two classes of next neighbor bins with common boundary lines or common corner points. A rectangular region of the variables x, y is therefore divided into basic hexagonal cells. The positions of the hexagons are shown in figure 8. A weight w_{ik} has to be assigned to every cell, which is proportional to its area. Hexagons in the interior of the histogram have weight 1, the corner cells $w = 1/4$, the other cells at the x minimum and maximum $w = 1/2$. Along the x axis and the parallel upper boundary, every second cell has $w = 1/2$.

A hexagonal histogram may be filled directly. Alternatively, a histogram with rectangular cells may be re-binned, computing the fractional overlaps between the rectangular and hexagonal cells.

As the substitute for the unknown rates r_{ik}, an auxiliary histogram v_{ik} has to be evaluated. Fractional intensity counting is done within rings of cells of rising rank ν around a given hexagon (i,k); the value $\nu = 1$ corresponds to the next neighbors. The maximum number of cells in a ring is 6ν and the total number of cells up to ring n is

$$n_{cell,ikn} = 1 + 3 \cdot n \cdot (n+1) \ . \tag{3.1}$$

Cells may lie partly outside the x and y boundaries. The effective number of bins in the two-dimensional case is the total cell weight

$$W_{ikn} = w_{ik} + \sum_{\nu=1}^{n} \sum_{m=1}^{6\nu} w_{jl} \, , \tag{3.2}$$

where

$$j(i,k,\nu,m) \qquad \text{and} \qquad l(i,k,\nu,m)$$

are the indices of the m-th neighbor hexagon. The third argument of the functions j, l is the rank ν of the neighbors. The averaged intensity of order n is defined as

$$v_{i,k} = \frac{h_{i,k} + \sum_{\nu=1}^{n} \sum_{m=1}^{6\nu} h_{j,l}}{W_{ikn}} \ . \tag{3.3}$$

To fix n, criterion (2.5) has to be fulfilled. One looks for the lowest integer n which gives

$$h_{i,k} + \sum_{\nu=1}^{n} \sum_{m=1}^{6\nu} h_{j,l} \geq (F \cdot \frac{W_{ikj}}{n_{cell,ikn}}) \cdot (\sum_{i} \sum_{k} h_{i,k}) \cdot (\frac{v_{i,k}}{v_{\max}})^{2\kappa} \tag{3.4}$$

with $v_{ik} > 0$. As for the one-dimensional case, the fraction F is reduced automatically, if the ring of cells is not fully contained in the allowed (x,y) region.

3.2 The algorithm

A hexagon in the interior of the (x,y) rectangle has 6 next neighbors. The generalization of the information exchange to two dimensions is

$$h^*_{ik} = h_{ik} \cdot (1 - \sum_{m=1}^{6} f_{ik,jl} \cdot \frac{1}{w_{ik}}) + \sum_{m=1}^{6} h_{jl} \cdot f_{ik,jl} \cdot \frac{1}{w_{jl}} \, , \tag{3.5}$$

where

$$j(i,k,1,m) \qquad \text{and} \qquad l(i,k,1,m)$$

are the indices of the m-th neighbor hexagon again.

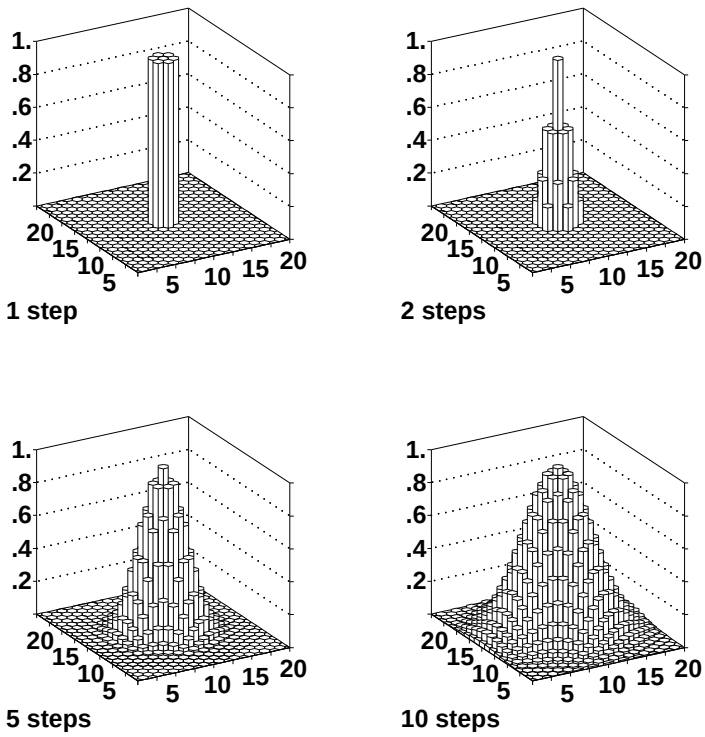

Figure 9: Two-dimensional diffusion of a bin content for some numbers of iterations.

The f coefficients are symmetric in the pairs of indices. The factors w_{ik}, w_{jl} take into account the lower cell areas at the boundaries: the mean histogram contents are proportional to w_{ik}, so that the density and the diffusion rate are proportional to h_{ik}/w_{ik}.

For the limiting case of shifting 6/7 of a cell content to the next neighbors one has $f = f_{\max} = 1/7$, which replaces the value 1/3 for the one-dimensional case. Some distributions, obtained from a single filled bin after several diffusion steps with $f = 1/7$, are shown in figure 9. The hexagonal symmetry and the falling intensity with increasing distance from the originally filled bin are obvious.

For fixed $f_{ik,jl} = f$ the variance of an original delta function becomes, after N diffusion steps,

$$\sigma_{ik}^2 \approx 6 \cdot f \cdot N \,, \tag{3.6}$$

measured in numbers of hexagon layers. In analogy to the one-dimensional case, this is set to the geometric variance of the cell system of rank n. If approximated by an elliptic region, this variance is

$$\sigma_{ik}^2 \approx \frac{1}{2} \cdot n \cdot (n+1) \,, \tag{3.7}$$

or, with eq. (3.1):

$$\sigma_{ik} \approx \sqrt{\frac{1}{6} \cdot (n_{cell,ikn} - 1)} \; . \tag{3.8}$$

Equations (3.8) and (3.6) can be combined with formula (2.5) to compute the number of diffusion steps:

$$N = \frac{1}{36 \cdot f_{\max}} \cdot F \cdot (\sum_i \sum_k h_{i,k}) \cdot (\frac{v_{\min}}{v_{\max}})^{2\kappa} \cdot \frac{1}{v_{\min}} \; . \tag{3.9}$$

The exchange coefficients are proportional to

$$f \sim v_{i,k}^{2\kappa-1} \, , \tag{3.10}$$

which leads to the ansatz

$$f_{ik,jl} = f_{\max} \cdot b(ik,jl) \cdot (\frac{2 \cdot v_{\min}}{v_{i,k} + v_{j,l}})^{1-2\kappa} \; . \tag{3.11}$$

The factor b is another complication introduced by different cell sizes: the exchange rate between two neighbor cells is proportional to the length of their common boundary line, which is 1/2 of the usual hexagon side length for a part of the cells at the histogram boundaries (see figure 8). In these exceptional cases one has $b = 1/2$ instead of the normal value $b = 1$.

The relations (3.3), (3.4), (3.11), (3.5) and (3.9) define the two-dimensional smoothing algorithm. From figure 8 one might get the impression that the procedure is based on Euclidian geometry. Actually this is not the case. For an infinitesimally fine binning, a rescaling of the x coordinate to $x \cdot \xi$ changes the histogram function from h to h/ξ and the number of bins has to be multiplied with ξ. According to eq. (3.11) the diffusion constants are invariant. The iteration formula (3.5) gives then a rescaling of h^* with the factor $1/\xi$. This internally consistent result is important, because in most physical applications the dimensions of the variables are different and a geometrical interpretation of the formalism is impossible. The number of steps N is modified, because it is proportional to the number of bins. This happens because now smearing over a different number of bins is needed to get the same fraction F of accumulated events.

The systematic errors in the two-dimensional case can depend on the direction of the diffusion. To keep this effect small, the widths of structures, measured in number of bins, should not differ very much in the x and y directions and the binning has to be chosen by the user accordingly.

The definition of the auxiliary spectrum v_{ik} is ambiguous. In analogy to the one-dimensional case, a self consistent approach can be introduced by using the same κ value in the pre-smoothing and the main analysis. Alternatively, a mixed approach can be constructed by using $\kappa = 0$ in the pre-smoothing step.

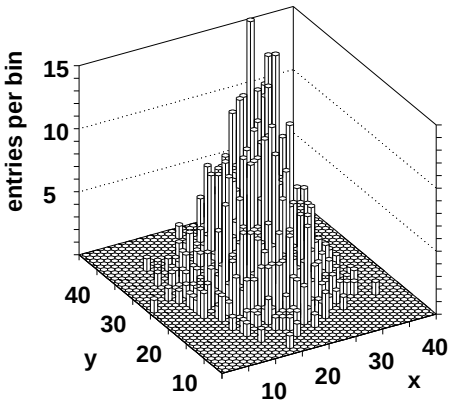

Figure 10: Two-dimensional histogram with 2000 Monte Carlo entries, generated from a rotational invariant Gaussian function with a width of 5 bins.

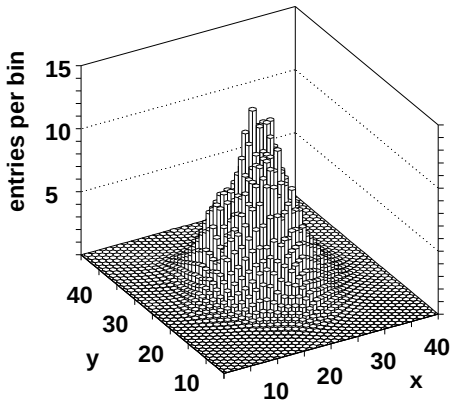

Figure 11: Smoothened histogram computed from the data of figure 10 with $F = 0.05$. Weak smoothing, mixed approach.

3.3 Implementation

The hexagonal histogram structure does not introduce much logistic overhead in an application program. The necessary tools exist and the 2-dimensional smoothing algorithm is very easy to use. There is a function to address a 2-dimensional array representing the hexagonal histogram, if the point coordinates are given. It is therefore trivial to fill a hexagonal histogram. It is possible to do a linear interpolation to an arbitrary point in the smoothened hexagonal histogram. As an interface to graphic representations, a routine exists which creates the graphics primitives necessary to draw hexagonal lego plots. An existing histogram with rectangular bins can be converted into the hexagonal format with a re-binning routine to allow smoothing. It is also possible to transform a histogram with hexagonal bins into a normal one. All technical details about pre-smoothing, weighting and iterations are hidden in the smoothing routines and a user has to call one interface routine only [8].

3.4 Systematic errors

All qualitative remarks made in section 2.4 apply here, too. The estimate of peak broadening is somewhat different. Repetition of the arguments of section 2.4 for strong smoothing with eqs. (3.9) and (3.11) gives

$$6 \cdot f \cdot N \approx 2 \cdot F \cdot \frac{n}{n_p + 4\pi \cdot \sigma_p^2 \cdot b} \cdot \sigma_p^2 \; . \tag{3.12}$$

Here, the fraction F enters linearly and not quadratically. The condition on the smallness of F is more restrictive in two dimensions; F should be $\leq 0.05 \cdot n_p/n$.

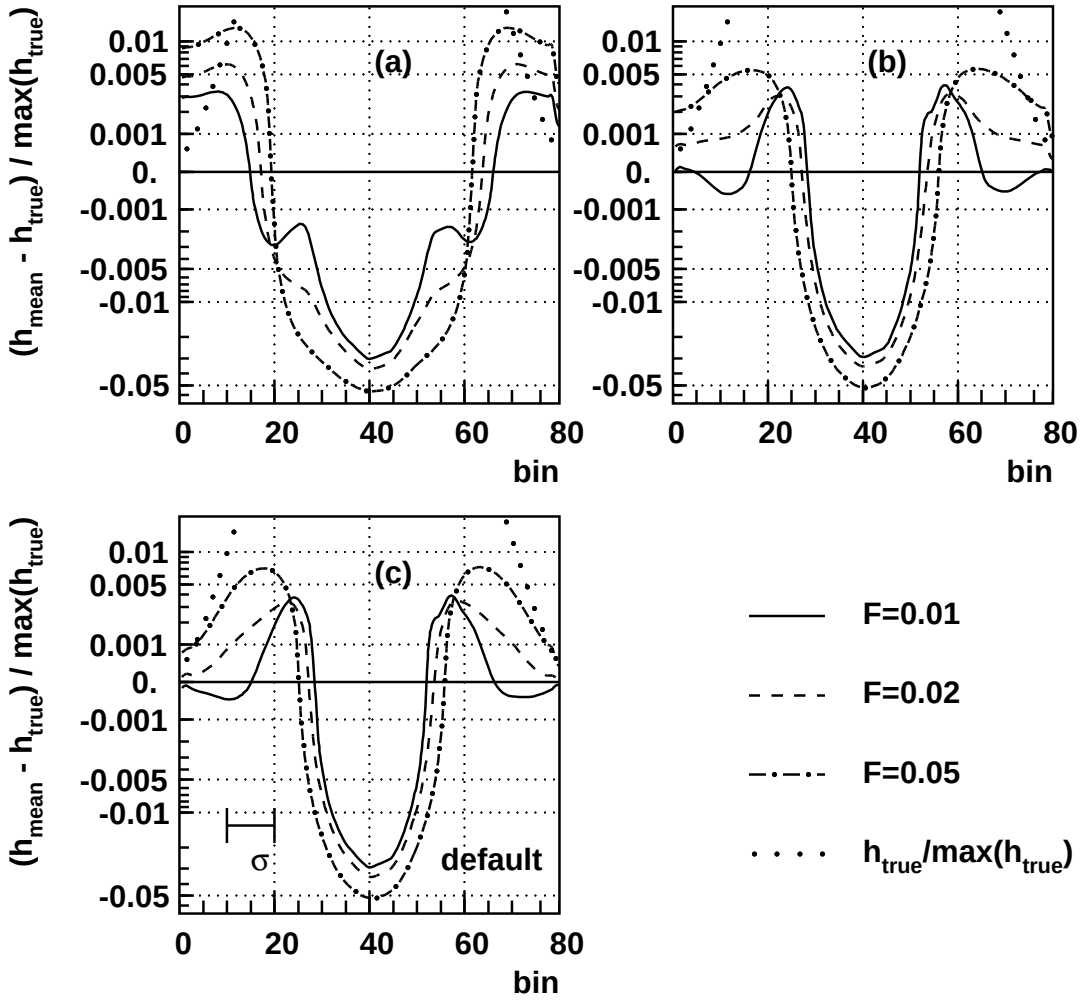

Figure 12: Deviation of the averaged smoothened 2-dimensional histograms h_{mean} from an original Gaussian function with an intensity of 2000 events and a variance of 10 bins. The differences are normalized to the true maximum. The results are shown for a band through the center parallel to the x axis, 10 bins wide. (a) strong smoothing, (b) weak smoothing, self consistent approach, (c) weak smoothing, mixed approach. Non-linear ordinate scale as in figure 4.

Gaussian function The study follows closely section 2.4. A Monte Carlo data set based on a 2-dimensional Gaussian is shown in figure 10, and one example for smoothing is given in figure 11. After smoothing one expects $\approx 1/F = 20$ statistical wiggles, superimposed on the Gaussian, which is indeed the case. They are all correlated to fluctuations in the input data set.

Cuts through averaged smoothened histograms, obtained with 10000 Monte Carlo data sets, each containing 2000 events, are presented in figure 12. To select the bands,

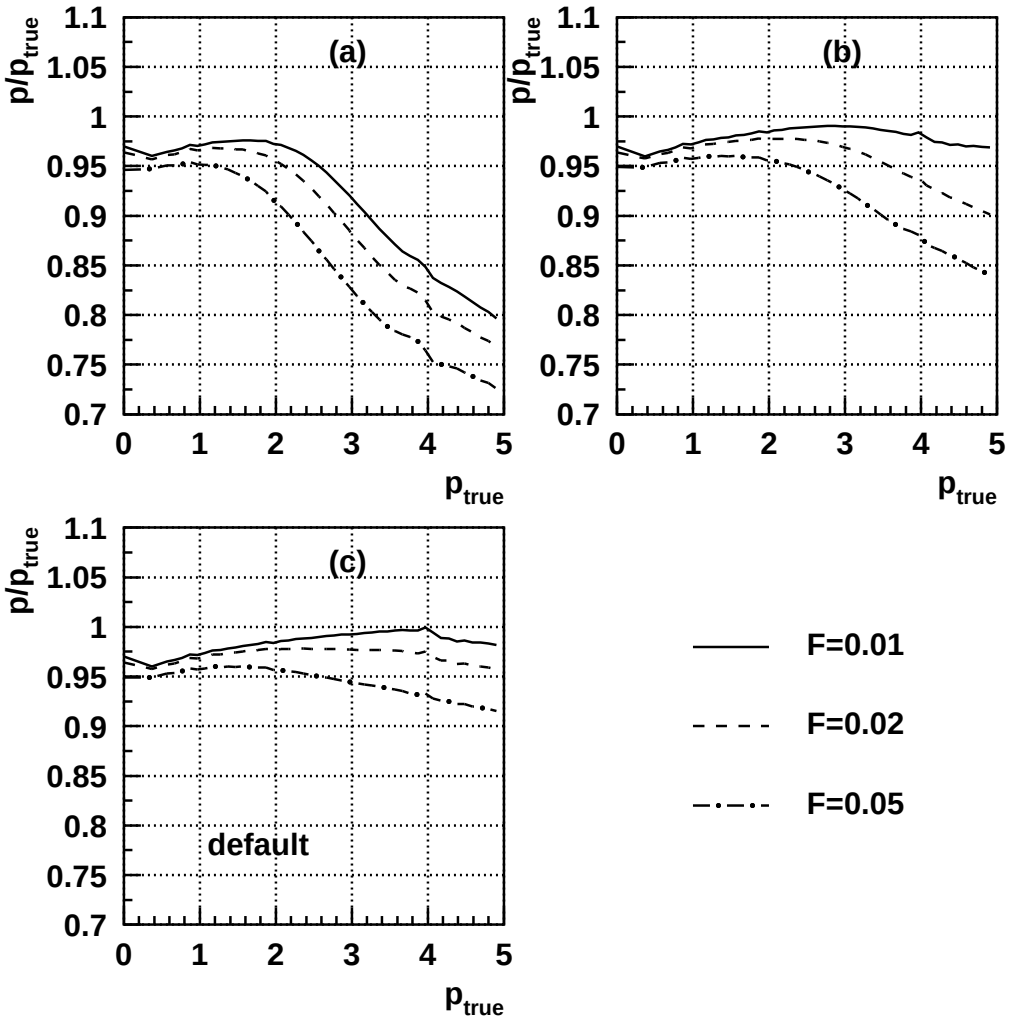

Figure 13: Differences between the mean reconstructed and true pulls p_{true} for a 2-dimensional Gaussian distribution as a function of the true pull. The Monte Carlo samples are the same as in figure 12. (a): strong smoothing, (b): weak smoothing, self consistent approach, (c): weak smoothing, mixed approach.

the hexagonal histograms were converted to the rectangular format. To avoid additional broadening of the distribution due to the rebinning, the number of bins and the variance were increased by a factor 2 with respect to to figures 10 and 11.

The cumulated averaged histograms were compared with the true Gaussian integrals. Like in section 2.4, equivalent Gaussian pulls are presented instead of the summed his-

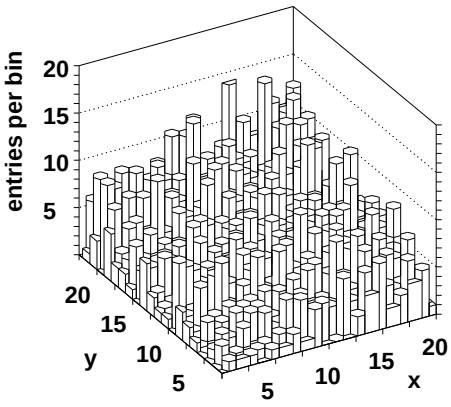

Figure 14: Two-dimensional histogram with 2000 Monte Carlo entries, generated from a linear function with a gradient along the x, y diagonal and vanishing rate at the origin.

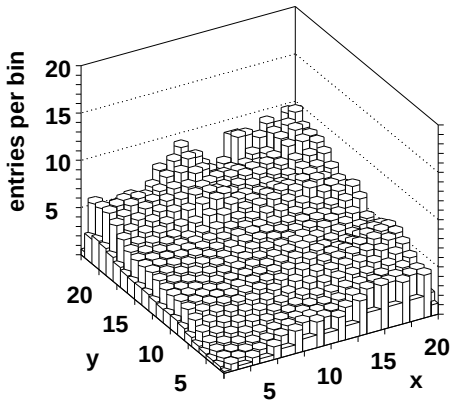

Figure 15: Smoothened histogram, computed from the data of figure 14 with $F = 0.05$. Weak smoothing, mixed approach.

togram contents. In two dimensions the pull is computed from

$$\frac{\sum_{h_{mean,m,n}/w_{mn} \le h_{mean,i,k}/w_{ik}} h_{mean,m,n}}{\sum_{m,n} h_{mean,m,n}} = 1 - \int_0^{p(i,k)} 2\xi \exp(-\xi^2) d\xi$$
$$= \exp(-p(i,k)^2) \ . \tag{3.13}$$

The center of the distribution corresponds to $p = 0$ and p extends to $+\infty$.

The pull ratios, representing the local widening of the distributions, are shown in the plots 13. Above the true pull $p_{\text{true}} = 4$ the distribution is partly truncated by the histogram boundaries, which gives sometimes rise to kinks, as visible in figure 13. As expected from the one-dimensional results, weak smoothing with the mixed approach ansatz is superior concerning the systematic shifts.

Planar function A numerical example is given in figures 14 and 15. The mean population density in the input histogram is 5 events per bin and the true maximum bin content is 10. This example is rather extreme, because the diffusion fills up mainly some corner bins close to the origin. The different bin weights at the histogram boundaries show up in figure 15, an effect which can be avoided by plotting h_{ik}/w_{ik}. Again statistical wiggles, residuals of the initial polynomial fluctuations, are observed. The most pronounced peak in the smoothened histogram arises from a data excess in four adjacent bins of the input at $x \approx 9, y \approx 19$.

Differences between the mean reconstructed and the true distributions are shown in figure 16. To construct cuts parallel to the coordinate axes, the smoothened histograms

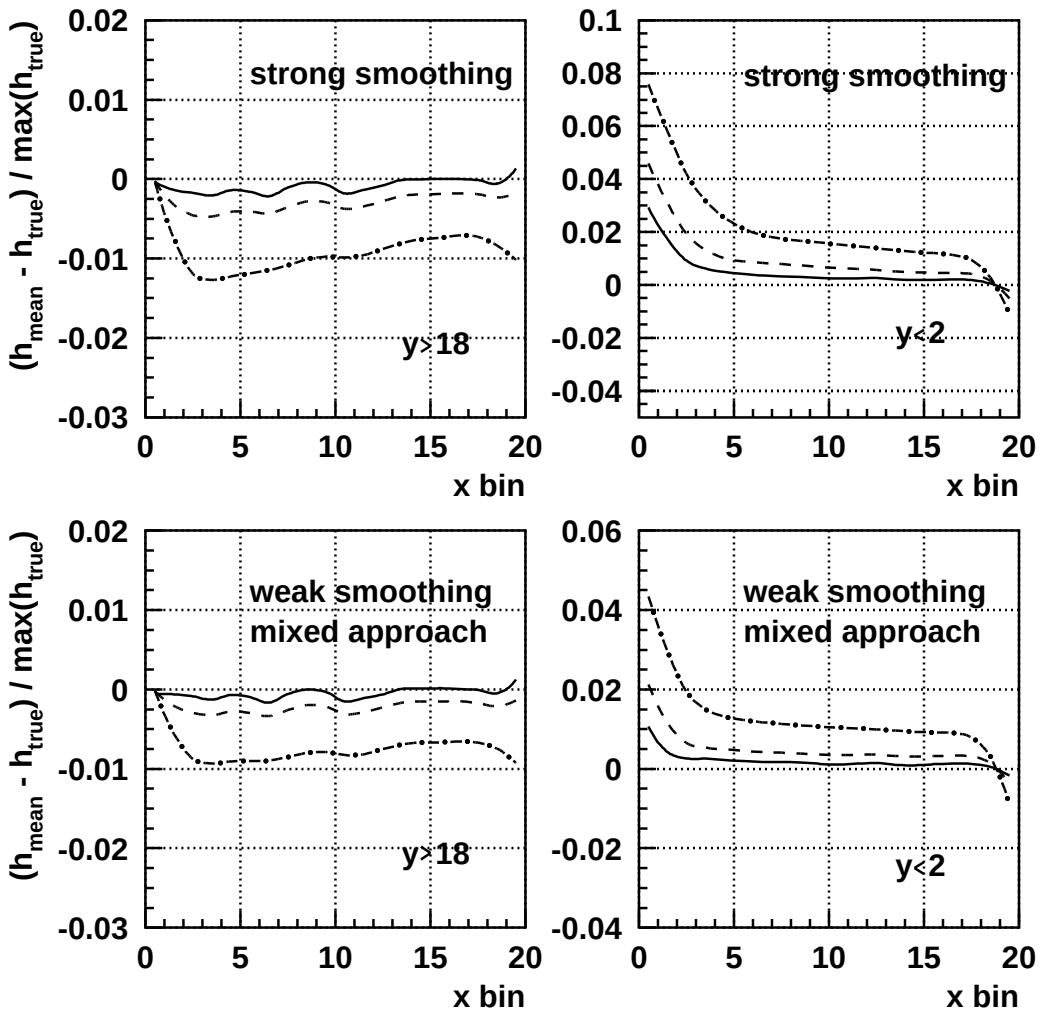

Figure 16: Deviation of the averaged smoothened histograms h_{mean} from the true 2-dimensional linear function of figure 14, normalized to the true maximum. The number of Monte Carlo data sets is 10000. Full lines: $F = 0.01$, dashed lines: $F = 0.02$, dash-dotted lines: $F = 0.05$.

were converted to the rectangular format. The rebinning introduces no extra systematic errors for a linear function and the true distribution and the binning are identical to that of figure 14. The reconstruction shows an event excess of the order of F at the origin $x = y = 0$ and a smaller excess at the coordinate axes. Along the parallel lines at $y = 20$ and $x = 20$ the reconstructed intensity is too low, the transition from the excess to the deficit appearing at the corner points.

Delta function In two dimensions, one starts with one filled hexagon as input. The

computation equivalent to section 2.4 gives with eq. (3.11) and (3.3)

$$f = \frac{2}{7} \cdot \frac{v_{\min}}{\max(v_{ik} + v_{jl})} \tag{3.14}$$

$$\max(v_{ik} + v_{jl}) = \sum_k h_k \cdot (1 + \frac{1}{7}) \; . \tag{3.15}$$

The tail intensity is

$$1 - (1 - 6f)^N \approx 6 \cdot N \cdot f = \frac{7}{24} F \; . \tag{3.16}$$

4. Conclusions

A diffusion like algorithm has been investigated for smoothing one- and two-dimensional histograms. Its two steering parameters F and κ are simply related to the statistical fluctuations after smoothing; they determine the overall size of the fluctuations and their dependence on the local rate.

The systematic errors depend on the steering parameters; they were evaluated for simple prototypes of distributions in detail. Any structure to be investigated should extend over several bins. The relative systematic errors are then insensitive to the binning and the number of entries in the histogram, if this number is sufficiently large. A compromise between smoothness of the result and systematic deformations of the distribution has to be found by the user.

In addition to the choice of steering parameters, there is an ambiguity in defining an initial approximation for the distribution. It has been found that best overall performance is obtained, if $\kappa = 0$ is used for the pre-smoothing instead of the κ parameter of the main analysis.

Compared to other available smoothing algorithms like multiquadric smoothing or spline fits, the diffusion algorithm produces less oscillations in regions of low intensity. With the default value $\kappa = 1/2$ the algorithm works in a similar way as the method of Gaussian kernels with a kernel width inversely proportional to the square root of the intensity. By shifting the second parameter κ towards lower values, fluctuations in low intensity regions can be reduced at the price of increased systematic errors such as peak broadening and peak tails. By shifting κ in opposite direction, the algorithm reaches finally the case of a coarser histogram binning, the fluctuations at low intensity become larger and the systematic errors are reduced. The mixed approach ansatz of weak smoothing, similar to the Gaussian kernel method, turns out to be a good compromise.

References

[1] A.L. Read, in *Proceedings of the Workshop on Advanced Statistical Techniques in Particle Physics*, Durham, 2002.

[2] *HBOOK Manual, CERN Program Library Y250*, CERN, Geneva, 1994.

[3] J. Allison, *Comput. Phys. Commun.* **77** (1993) 377.

[4] K.S. Cranmer, *Comput. Phys. Commun.* **136** (2001) 198.

[5] W. Murray and V. Ruhlmann-Kleider, *Estimation of probability density functions for the Higgs search*, *DELPHI public note* **2000-067** PROG 240, `http://delphiwww.cern.ch/pubxx/delnote/dn2000.html`.

[6] I. Abramson, *Ann. Statist.* **10** (1982) 1217.

[7] OPAL collaboration, G. Abbiendi et al., *Search for the standard model Higgs boson with the OPAL detector at LEP*, *Eur. Phys. J.* **C 26** (2003) 479 [`hep-ex/0209078`].

[8] The code is available at `http://www.physi.uni-heidelberg.de/~bock`.

JHEP08(2006)056