Superresolution algorithms for data analysis of discrete detectors in nuclear physics

E.A.Kolganova [a], E.L.Kosarev [b], G.A.Ososkov [a]

[a] *Joint Institute for Nuclear Research, Dubna, 141980 Moscow Region, Russia*
e-mail: `ososkov@cv.jinr.ru`

[b] *P.L.Kapitza Institute for Physical Problems, Russian Academy of Sciences, ul.Kosygina, 2, Moscow 117334, Russia*
e-mail: `kosarev@kapitza.ras.ru`

Abstract

It is shown that the recovery of particle coordinates by detectors with intrinsic resolution determined by the point spread function and finite size of detector bins can be reduced to a solution of the standard convolution integral equation with a modified point spread function. Two approaches are proposed and investigated for this problem: parametric and non-parametric ones. Algorithms and their testing for both approaches are given. It was shown that both algorithms can resolve the coordinates of particles with resolving power better than the bin-size of the detector granulation unit and the point spread function characteristic scale. It is also demonstrated that the superresolution efficiency of the proposed parametric algorithm almost attains the Cramér-Rao limit and Shannon's limit for the non-parametric algorithm. Results of numerical experiments and of real data processing of the CERES silicon drift detector are given.

PACS: 02.50.Rj; 02.60.Ed; 07.05.Kf; 29.85. +c

Keywords: Superresolution; Histograms; Silicon Drift Detector; Detector Granularity; Accuracy; Efficiency; Algorithm; Maximum Likelihood.

1 Introduction

This paper is devoted to the problem of the most exact determination of the observed signal position from noisy experimental data distorted by a measuring device and written as histogram data. The problem has a long history [1]-[3] and is of importance for data processing in almost any field of experimental physics: in time-of-flight techniques, spectrometry, astronomical imaging, processing of signals from silicon drift detectors (SDD) or time projection chambers etc [4]-[11]. Therefore our consideration is mainly of a methodological character,

[1] A short version of this paper was presented as a report to the International Conference "Computing in High Energy Physics", Berlin, Germany, April 1997

which gives the reader a general criterion for choosing a data handling approach by comparing its accuracy with an ideal, the best possible one. Our methods described below are general enough to be applicable in many specific experimental situations.

For better understanding we use the typical example of an elementary particle that passes a discrete detector producing an electromagnetic shower. In order to improve the resolution and accuracy of contemporary detectors listed above, they are designed as granular structures consisting of an array of cells (pads). During the process of the registration the signal is spread between several adjacent cells, so each cell registers a portion of electron shower energy or charge fallen on it. Thus the signal during the registration is discretized and stored as a 1D or 2D histogram.

The basic problem is: to reconstruct the original signal position and its other parameters (its amplitude or the volume under its surface, its half-width etc) from the registered histogram. Depending on its formulation this problem can be solved in either non-parametric (NP) or parametric ways.

The first non-parametric approach also named *the unfolding problem* is applied when the parametrization of the problem is unknown. Mathematically it is formulated as the integral equation of the first kind (see Eq.(2) in the next section).

The second approach is well-known as parameter fitting. It is, usually, carried out by the least squares method (LSM). However, one of the most widely used ways for finding the position of a signal is based on the assumption of signal symmetry. It leads to the center of gravity (COG) algorithm. More accurate methods apply more detailed parametrization based on the approximation of the bell-shape form of the electron cloud by a two-dimensional gaussian or other symmetrical surface (see, for example [4,5,13]).

However for almost any real experimental set-up both problems are considerably complicated by, at least, three main factors:

- the signal distortion due to the histogramming process of the experimental data storing;
- the presence of background noise;
- signal overlapping due to the high occupancy of the majority of modern experimental systems.

The influence of the first factor, as it is shown in Section 2, can be expressed mathematically by the convolution of the so-called instrumental function of our detector (which is the kernel of the integral equation mentioned above) with the function (3) in the Section 2.1 presenting the finite size of the detector granularity.

The second factor develops in two ways: in adding some noise to each histogram value and in the appearance of some noisy signals exceeding a given cut-off. The latter effect can be handled by the cut-off increase. New efficient robust methods are developed [15] and can be applied to process data contaminated by this type of background.

However noise distortions of the signal value registered in each detector pad can hinder the recovery of close signals. In fact these noise distorsions set the resolution limit for close signals. It follows from the close analogy between processing of noisy experimental data and data transmitted through a noisy channel considered in the next section.

The third factor which is very important in many cases is when the probability of two and more overlapping signals is sufficiently large, so this paper is focused on it in its main part. We do not consider here the wavelet transform approach developed recently for extracting of parameters of noisy signals [12], since it is definitely amenable to the parametric methods proposed below, especially in the resolution accuracy of very close signals.

The paper is organized as follows. After the present introduction the basic mathematical background is given in section 2 devoted to the non-parametric approach. Then the problem of taking account of both the intrinsic resolution of the detector and the finite size of detector cells is considered. It is reduced to the standard convolution type equation with the modified kernel shape. It is also explained how the well-known Shannon theorem [14] can be applied to find a limit of the reachable accuracy of signal resolution. The non-parametric maximum-likelihood algorithm for recovery of particle data is also described in the end of this section.

Section 3 shows how the more accurate parametric estimation (PAM - see further) can be applied to the problem and the Cramér-Rao (CR) lower-bound expression for the accuracy limit of particle location coordinate x_0 is derived. In section 4 numerical results are presented of 1D simulations as well as of some examples of real SDD data processing. Results of NP-PAM applications are compared with COG-method applications and the corresponding accuracy of CR lower-bound. Concluding remarks are presented in section 5.

2 Non-parametric approach

For the sake of intelligibility let us use the CERES SDD doublet example with its polar co-ordinates (although for an cartesian co-ordinate SDD system like wafers of STAR-SVT (Silicon Vertex Tracker) set-up [11] the problem formulation is almost the same as in [16]). Each of two CERES SDD detectors has a disk shape with the active inner and outer radii of $R_{in} = 6\ mm$ and $R_{out} = 32\ mm$, respectively (see Fig.1).

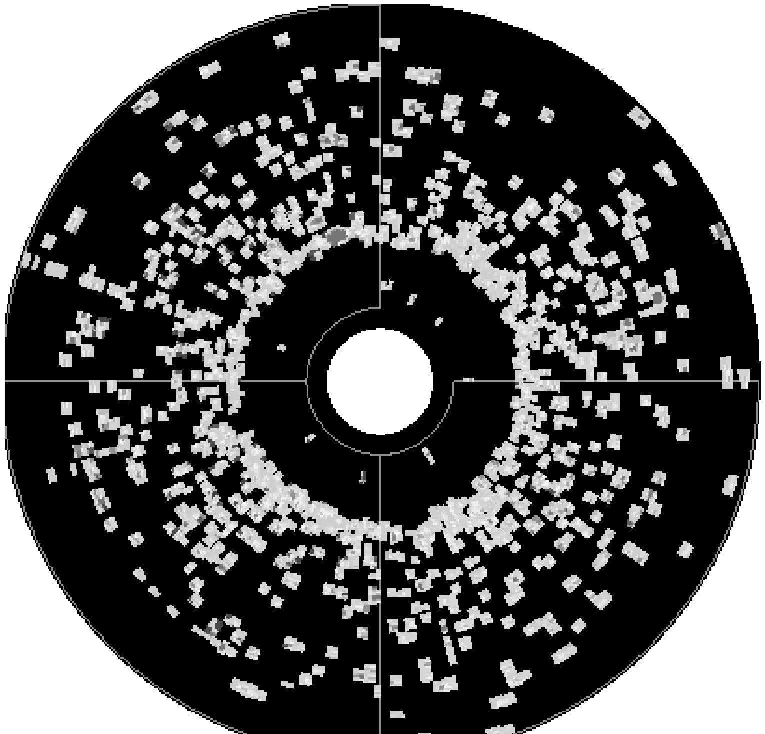

Fig. 1. CERES SDD image of an Au-Pb event.

The 360 radially oriented anodes register signals, i.e. the charges of electron clouds appearing when particles hit the detector. While drifting an electron cloud increases its size due to diffusion and electrostatic interaction between electrons.

As it was shown in [5] the signal has a bell-shape form, which can be reasonably be approximated by a two-dimensional Gaussian with maximum amplitude A:

$$N(x, r; A, x_0, R_0) = A \exp\left[-\frac{(x - x_0)^2}{2{\sigma_x}^2} - \frac{(r - R_0)^2}{2{\sigma_r}^2} \right], \tag{1}$$

where $x = r\varphi$ is perpendicular to the radius direction, R_0 and $x_0 = R_0\varphi_0$ are the initial coordinates of the electron cloud, and

$${\sigma_r}^2 = 2Dt, \quad {\sigma_x}^2 = 2Dt\frac{R_{out}}{R_0}$$

are their variances,

$$t = \frac{R_{out} - R_0}{V_d}$$

is the drift time, R_{out} is the radius of the electron cloud at the time t. Values D for the diffusion constant of electrons in silicon, and the radial drift velocity V_d are supposed to be known constants.

The single particle event (electron cloud charge) is registered by several adjacent cells of a 2D grid formed by 360 anodes in the azimuthal direction and time-bins in the radius direction. So the charge distribution on each cell can be calculated by 2D integration of (1) over this cell. If two signals ovelap, their contributions to the same cell are superimposed.

Our aim is to estimate the center and the amplitude of a digitized signal, which can be considered as two-dimensional histogram $\{a_{ij}\}$ formed by extracting a cluster of ajacent cells with amplitudes exceeding a given threshold.
However, due to the factorized form of a 2D Gaussian (1) we can reduce this 2D problem to several one-dimensional ones. As it can be seen in Fig.1 due to an extremely high multiplicity of central Au-Pb collisions (up to 1000 tracks) over 30% of signals overlap each other. If signals occur so close to each other that they are indistinguishable producing only one maximum, a correct parametrization is hard to obtain, since even the number of functions to be fitted is unknown. However the non-paramentric algorithm described below can be used to produce good initial parameter values for the parametric scheme based on reducing the problem to the case of handling of one-dimensional histograms.

2.1 Problem formulation

According to the number of anodes covered by the given cluster (let it be k) we split the 2D array $\{a_{ij}\}$ for each fixed φ_j into k one-dimensional histograms $\{F_i\}$ for $i = 1, 2, ..., n_k$. Then for each j one can obtain the function $f(x)$ as a solution of the following integral equation

$$F_i = \int\limits_{\Delta_i} dr \int K(r - x) f(x) dx = \int\limits_{r_i - \Delta/2}^{r_i + \Delta/2} dr \int K(r - x) f(x) dx, \tag{2}$$

where the internal integral in (2) is taken with respect to x, the function $f(x) > 0$ is unknown and the kernel $K(r)$ is the one-dimensional gaussian

$$K(r) = \exp\left(-\frac{r^2}{2{\sigma_r}^2}\right)$$

with σ_r supposed to be a known constant inside the given cluster. To take into account the finite size of our histogram bin we determine the function

$$h(r) = \begin{cases} 1, \ |r| \leq \Delta/2 \\ 0, \ |r| > \Delta/2, \end{cases} \tag{3}$$

and transform the formula (2) as follows:

$$F_i = F(r_i) = \int\limits_{r_i-\Delta/2}^{r_i+\Delta/2} dr \int K(r-x)f(x)dx = \int\limits_{r_i-\Delta/2}^{r_i+\Delta/2} h(r-r_i)dr \int K(r-x)f(x)dx =$$

$$= \int f(x)dx \int\limits_{r_i-\Delta/2}^{r_i+\Delta/2} h(r-r_i)K(r-x)dr. \tag{4}$$

Then substituting the integration variable in the internal integral in (4) by $r - x = t$ and using the fact that the function $h(r)$ is even we obtain

$$\int h(t+x-r_i)K(t)dt = \int h(r_i - x - t)K(t)dt = K_1(r_i - x). \tag{5}$$

Here the modified instrumental function

$$K_1(s) = \int h(s-r)K(r)dr \tag{6}$$

is the convolution of the function h and the original kernel K. We obtain eventually the following integral equation

$$F_i = F(r_i) = \int K_1(r_i - x)f(x)dx. \tag{7}$$

Since our measurements are, in fact, the values F_i of the histogram, whose bin centers are denoted as x_i, the integral equation (7) is reduced to the system of algebraic equations suitable for computer implementation.

$$\sum_{j=1}^{m} P_{ij}G_j = F_i, \ i = 1, 2, ...n, \tag{8}$$

where

$$P_{ij} = K_1(r_i, x_j) \cdot (x_{j+1} - x_j) \tag{9}$$

Depending on its formulation this problem can be solved in either non-parametric or parametric ways. The first non-parametric approach also named *the unfolding problem* is applied when the problem parametrization is unknown. Mathematically it is formulated as the integral equation of the first kind (2) with the kernel (6).

Although for continuous kernels the unfolding problem is ill-posed (see [17]), however as it was proved there, when one looks for a solution on a compact set, the problem can be solved for a sufficiently wide class of kernels including the convolution $K_1(s)$.

2.2 Maximum-likelihood (ML) algorithm for convolution integral equations with the Gaussian distribution of input data

The contents of this section is based on [18] and [19], where readers can find more detailed information on this subject. In the case of a Gaussian distribution function of input data the logarithmic likelihood function may be written as

$$L = -\frac{1}{2}\sum_{i=1}^{n}\frac{(F_i - S_i)^2}{D_i} + const, \tag{10}$$

where the values of D_i are equal to the noise variances at the ith experimental point

$$D_i = \sigma_i^2, \quad i = 1, 2, ..., n\,,$$

and the values of S_i are defined by the formula

$$S_i = \sum_{k=1}^{m} P_{ik}G_k, \quad i = 1, 2, ..., n.$$

with P_{ik} taken from (9). Let us denote the sum of G_k by G and the normalized quantities g_k

$$G = \sum_{k=1}^{m} G_k, \quad g_k = \frac{G_k}{G},$$

so the likelihood function may be defined in the extended $(m+1)$-dimensional space $\{g_k, G\}$, $k = 1, 2, ..., m$. The extremum conditions in this space are as follows

$$\frac{\partial L}{\partial g_k} = G\sum_{i=1}^{n}\frac{F_i - S_i}{D_i}P_{ik} = 0, \quad k = 1, 2, ..., m \tag{11}$$

and

$$\frac{\partial L}{\partial G} = G^{-1}\sum_{i=1}^{n}\frac{F_i - S_i}{D_i}S_i = 0\,. \tag{12}$$

From the (11) one may obtain

$$G = \frac{\sum_{i=1}^{n}(F_i s_i)/D_i}{\sum_{i=1}^{n} s_i^2/D_i}, \quad \text{where } s_i = \sum_{k=1}^{m} P_{ik}g_k = S_i/G.$$

So for any given vector $\{g_k\}$ there is only one value of G maximizing the likelihood function.

The maximum of the likelihood function L is sought for the case of Gaussian noise by means of an iterative procedure. The direction of search $\delta g_k^{(t)}$ is defined by

$$g_k{}^{(t+1)} = g_k^{(t)} + h\delta g_k^{(t)}, \text{ where } \delta g_k^{(t)} = g_k^{(t)} \sum_{i=1}^{n} \left(\frac{F_i - G \sum_{j=1}^{m} P_{ij} g_j^{(t)}}{D_i} \right). \tag{13}$$

The value of the optimal step h in the direction of search $\{\delta g_k\}$ is defined from the equation

$$\frac{d}{dh} L(G(h), \{g_k + h\delta g_k\}) = 0, \tag{14}$$

where

$$G(h) = \frac{\sum_{i=1}^{n} (F_i s_i)/D_i + h \sum_{i=1}^{n} (F_i \delta s_i)/D_i}{\sum_{i=1}^{n} (s_i + h\delta s_i)^2/D_i},$$

and

$$\delta s_i = \sum_{k=1}^{m} P_{ik} \delta g_k.$$

It was shown in [19] that the likelihood function after one iteration step

$$L(h) - L(0) = \frac{1}{2} G(h) \sum_{k=1}^{m} \frac{(\delta g_k)^2}{g_k} h \geq 0$$

is always non-negative and proportional to the length squared of the likelihood function gradient, so the iteration process (13) with step h defined by (14) and satisfying the condition

$$h \leq \min_{\{k; \delta g_k < 0\}} (-g_k/\delta g_k)$$

converges to the maximum of the likelihood function (10).

The iteration formula (13) is essentially nonlinear, firstly, because the unknown vector $\{g_k\}$ is used here as the factor and, secondly, due to the dependence of s_i on g_k. The existence of such nonlinearity in the ML-algorithm is its most important property, and it is the key to the ultimate resolution achievement as compared with any linear methods of signal recovery. It is shown in [18] that this improvement of resolution, or superresolution, by using (13) reaches the theoretical limit. Such a limit for close signals follows from the well-known Shannon theorem for the maximum speed of data transmission via a channel containing noise[14].

To explain that in more detail let us introduce some useful definitions concerning the super-resolution factor. The **resolution** of any linear device with the instrumental function $K(x)$ can be defined as the effective width of this function, i.e.

$$D = \int_{-\infty}^{\infty} K^2(x) dx, \tag{15}$$

providing the normalizing condition at the origin $K(0) = 1$.

The resolution of spectral devices can be improved in comparison to D using modern techniques for solving integral equations, thus superresolution is achieved. We define **the superresolution factor** as the ratio of D to the separation δ between two narrow lines which can be recovered after the deconvolution procedure

$$SR = D/\delta \; . \tag{16}$$

We should recall that according to the well-known Rayleigh definition of resolution $\delta = D$, so in this case the superresolution factor is $SR = 1$. When the resolution is improved mathematically, $SR > 1$. The improvement is always limited by noise. At zero noise an exact solution of Eq.(2) can be found, which corresponds to an infinite superresolution.

The highest possible superresolution factor is closely related to Shannon's theorem on the highest possible transmission rate of information through a noisy channel [14]. When a spectrum is not parametric, i.e. the function we sought for cannot be described by a simple formula with a few parameters, the limiting superresolution factor is

$$SR = \frac{1}{3}\log_2(1 + E_s/E_n). \tag{17}$$

Here E_s and E_n

$$E_s = \int_{-\infty}^{\infty} F^2(x)dx, \; E_n = n\sigma^2,$$

are the signal energy and the noise one. Here n is the number of experimental data points, and σ^2 is the variance of input noise. If the signal-to-noise ratio is expressed in decibels $dB = 10\log(E_s/E_n)$, the approximate expression for the superresolution limit is

$$SR \simeq dB/10. \tag{18}$$

Let the signal-to-noise-ratio be for example 30 dB. It follows from Eq.18 that $SR \simeq 3$. This means that in this case we can resolve some details in the reconstructed signal on a minimal distance 3 times smaller than the size D of the instrumental function.

When the signal sought for can be described by a formula with a few parameters, the superresolution is determined by the Cramér-Rao lower-bound inequality [20] and it may be much higher than the one predicted by Eq.17. But when the shape of the spectrum is unknown and the aim of the experiment is to determine the shape, the parametric approach does not work, and resolution is determined by Eq.17.

In [19] the computer program package RECOVERY was described, which can reach the Shannon superresolution limit. The programs from the RECOVERY package are based on the maximum likelihood principle and they look for the maximum of the likelihood function in the finite-dimensional set of solutions, which is always a compact set. The RECOVERY code is accessible from the CPC Program Library. The results of the non-parametric approach described below in section 4 are obtained by using the modified DCONV program from the RECOVERY package. From now on we refer to the non-parametric method as NP.

3 Parametric approach and Cramér-Rao lower-bound for accuracy

It is obvious that parametric methods must be more accurate than non-parametric ones, since the parametrization itself brings essential information related to the processed signals. Even a general knowledge such as the signal symmetry is enough to apply the easy-to-calculate COG method to estimate the signal centroid:

$$x_{cog} = \frac{\sum_i a_{i,j}\bar{x}_i}{\sum_i a_{i,j}} \qquad y_{cog} = \frac{\sum_j a_{i,j}\bar{y}_j}{\sum_j a_{i,j}}, \tag{19}$$

where $a_{i,j}$ is a 2D histogram presenting a detector response to the current signal, $\bar{x}_i, \bar{y}_j$ are the middle points of the corresponding bins. The high speed and universality of this method makes it the most popular for the majority of discrete detectors, unless the growing particle multiplicity leads to the situation when 30% and more of signals overlap each other as one can see in Fig.1. Signal superposition causes considerable errors in the COG method despite various tricks with testing if a histogram is non-unimodal and then splitting it up into two or more clusters. Moreover, cases when signals appear so close that the resulting histogram becomes unimodal lead to the complete breakdown of the COG approach. The NP methods described above are rather slow and they do not satisfy the requirements of the very high rate of the data stream. Thus more elaborate parametric methods are needed to develop fast, efficient and accurate algorithms which could meet these requirements. Due to non-linearity of the problem these methods are necessarily iterative, so their quality depends very much on the right choice of initial parameter values. Reasonable sources of these initial values are either COG methods, when either the signal number is known or else results of the few first iterations of NP methods.

One more detail should be pointed out to specify requirements of the parametric methods in nuclear physics. Recent measurements on the CERES experiment show that the integrated front-end electronics violate the assumption about the full symmetry of the registered signals (see, for example, [13]). Namely for each radius R the values of σ_x are different. During a special statistical processing of Pb-Pb experimental data it was found that the appearing asymmetry can be locally described as a so-called "asymmetrical gaussian". This means that two halves of the signal pulse are fitted using a composition of two gaussians with widths $\sigma_0 - \Delta$ below the maximum, and $\sigma_0 + \Delta$ above the maximum:

$$f(x) = \frac{A}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right],$$

where $x \equiv x_i$ are time bins, A is amplitude, x_0 is a position of the gaussian maximum, and

$$\sigma = \begin{cases} \sigma_0 - \Delta & for \quad x \leq x_0, \\ \sigma_0 + \Delta & for \quad x > x_0. \end{cases} \tag{20}$$

The corresponding parameters were fitted on the representative sample of central Pb-Pb collisions, where only a small admixture of double pulses was observed. It yielded $\sigma_0 = 1.73$ time bins at $x_0 = 50$ time bins and $\sigma_0 = 2.09$ time bins at $x_0 = 220$ time bins [13]. The dependence is rather linear, namely $\sigma_0 = 1.626 + 0.2097 \cdot 10^{-2} x_0$. The delta parameter drops, $\Delta = 0.18$ at $x_0 = 50$, and $\Delta = 0.02$ at $x_0 = 220$.

This asymmetry observation is also an additional objection against an uncritical use of COG methods.

However, in our first approach we simplify the situation assuming that signals are symmetrical. Furthermore in section 4 we extend our formalism to the case of real signals with their asymmetry. We take into account the described asymmetry model in our simulations by applying σ from (20) in corresponding ranges of x.

3.1 Single peak parameters estimation

In the case of one single signal peak with the gaussian shape

$$\psi(r; A, R) = A \exp\left[-\frac{(r-R)^2}{2{\sigma_r}^2}\right] \tag{21}$$

we have the histogram $\{a_i\}, i = \overline{1,n}$. Let us suppose bin width to be unity, i.e. $\Delta r = r_{i+1} - r_i = 1$. We have to fit the function given by (21) to this histogram. Assuming that the noise distribution is normal with zero mean and known variance σ^2_{noise} the problem is reduced to minimising the corresponding least squares functional

$$\mathcal{L}(A, r_0) = \sum_i \left(a_i - \int\limits_{r_i}^{r_{i+1}} \psi(r; A, r_0) dr \right)^2 \tag{22}$$

It contains the unknown parameter r_0 under the integral sign. This obstacle can be avoided by replacing each integral in (22) by its approximate mean value: $\psi(\bar{r}_i; A, r_0)$, where $\bar{r}_i = (r_{i+1} + r_i)/2$, so (22) is simplified to

$$\tilde{\mathcal{L}}(A, r_0) = \sum_i \left[a_i - \psi(\bar{r}_i; A, r_0)\right]^2 \tag{23}$$

Searching for its minimum one should solve the corresponding system of normal equations obtained by equating to zero the $\tilde{\mathcal{L}}(A, r_0)$ partial derivatives. However, this system is, unfortunately, transcendental. This requires us to develop a special iterative procedure to solve it. As initial values $A^{(0)}, r_0^{(0)}$ of unknown parameters for this procedure in a single signal case we use

$$A^{(0)} = \max_{\{\hat{r}_i\}} a_i$$

and the center of gravity obtained from (19)

$$r_0 = \frac{\sum\limits_i a_i \bar{r}_i}{\sum\limits_i a_i}.$$

Then considering (23) as a function of two parameters $z = z(x, y)$ we approximate it in the vicinity of $A^{(0)}, r_0^{(0)}$ by an elliptic paraboloid

$$z = ax^2 + by^2 + cxy + dx + ey + f, \tag{24}$$

where x, y are current values of parameters r, A. To find six coefficients of (24) it is necessary to calculate the values of $\tilde{\mathcal{L}}(A, r_0)$ in the point template, i.e. in six specially selected points surrounding $x^0 = r_0^{(0)}, y^0 = A^{(0)}$ chosen as the base point of this template. We use the simplest template design: the base itself, step left, step right from it in each dimension and the last point by step right in both dimensions. After solving the corresponding system of six linear equations to find our paraboloid coefficients its minimum coordinates are easily calculated

$$x_{min} = \frac{-2bd + ce}{4ab - c^2}, \quad y_{min} = \frac{-2ae + cd}{4ab - c^2} \tag{25}$$

The obtained coordinates are used as $A^{(1)}, r_0^{(1)}$, i.e. as the base point for the second iteration. Although a few iterations are enough, as a rule, the iteration process is controlled by testing either the maximum admissible accuracy of the minimum position or the fixed maximal number of iterations. From now on we refer to this method as to Paraboloidal Approximation Method - PAM.

3.2 Cramér-Rao accuracy limit

For the sake of simplicity let us consider a model of estimating of the single signal location parameter x_0 from a sample of n measurements:

$$y_i = A\psi(x_i - x_0) + \varepsilon_i, \qquad i = 1, 2, .., n.$$

Here $\psi(x)$ is the signal shape, A is its amplitude, and a random noise ε_i has the Gaussian distribution $N(0, \sigma)$, i.e.

$$p(\varepsilon_i)d\varepsilon_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) d\varepsilon_i \ ,$$

where different noise sample units are assumed to be (for a simplicity again) independent and non-correlated: $\overline{\varepsilon_i \varepsilon_j} = \sigma^2 \delta_{ij}$. Applying the standard maximum likelihood method (MLM) for estimating the signal amplitude A and location parameter x_0 one has to maximize the likelihood function of the sample

$$L = \prod_{i=1}^{n} p(\varepsilon_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right) =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - A\psi(x_i - x_0)]^2\right].$$

This reduces to maximizing the logarithmic likelihood function

$$l = \ln L = Const - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - A\psi(x_i - x_0)]^2 \ .$$

According to the Cramér-Rao inequality [20] *for unbiased estimates* one has

$$D(x_0) \geq \frac{1}{E\left[\left(\frac{\partial l}{\partial x_0}\right)^2\right]} \; .$$

Here, as usual, the symbol E denotes the mathematical expectation obtained by various random realizations of ε_i for $i = 1, 2, ..., n$. Taking the partial derivatives with respect to the parameter x_0 one obtains

$$\frac{\partial l}{\partial x_0} = \frac{A}{\sigma^2} \sum_{i=1}^{n} \varepsilon_i \frac{\partial \psi(x_i - x_0)}{\partial x_0} \; , \quad \left(\frac{\partial l}{\partial x_0}\right)^2 = \frac{A^2}{\sigma^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \psi_i}{\partial x_0} \frac{\partial \psi_j}{\partial x_0} \varepsilon_i \varepsilon_j \; ,$$

where

$$\frac{\partial \psi_i}{\partial x_0} = \frac{\partial \psi(x_i - x_0)}{\partial x_0} \; .$$

Since $\overline{\varepsilon_i \varepsilon_j} = \sigma^2 \delta_{ij}$, one has

$$E\left[\left(\frac{\partial l}{\partial x_0}\right)^2\right] = \frac{A^2}{\sigma^4} \sum_{i=1}^{n} \sigma^2 \left(\frac{\partial \psi_i}{\partial x_0}\right)^2 = \frac{A^2}{\sigma^2} \sum_{i=1}^{n} \left(\frac{\partial \psi_i}{\partial x_0}\right)^2 \; .$$

As a result of these rather tedious calculations, we obtain the following formula for the Cramér-Rao lower-bound for the accuracy of the signal location parameter x_0 under condition of negligible correlation between x_0 and A

$$D(x_0) \geq \frac{\sigma^2}{A^2} \cdot \frac{1}{\sum_{i=1}^{n} \left[\frac{\partial \psi(x_i - x_0)}{\partial x_0}\right]^2} \; . \tag{26}$$

This final formula is valid for arbitrary shapes of the signal $\psi(x)$ and arbitrary ratio between the characteristic scale D of the function $\psi(x)$ and the bin width Δ (the latter can be treated as the distance between measured points).

Let us consider two cases. In the first one when $\Delta \ll D$, the Eq.(26) can be approximated by substituting the sum in the denominator by the corresponding integral.

$$\Delta \cdot \sum_{i=1}^{n} \left[\frac{\partial \psi(x_i - x_0)}{\partial x_0}\right]^2 \approx \int_{-\infty}^{\infty} \left(\frac{\partial \psi}{\partial x}\right)^2 dx. \tag{27}$$

Here Δ denotes the histogram bin width. Computing the corresponding integral for the signal shape $\psi(x) = \exp(-(x/D)^2)$ one obtains eventually the formula

$$D(x_0) \geq \frac{\sigma^2}{A^2} \cdot \frac{1}{\sum_{i=1}^{n} \left[\frac{\partial \psi_i}{\partial x_0}\right]^2} \approx \frac{\sigma^2}{A^2} \cdot \frac{D\Delta}{\sqrt{\frac{\pi}{2}}}$$

and its approximate estimation

$$\delta x_0 = \sqrt{D(x_0)} \approx \frac{\sigma}{A} \cdot \frac{\sqrt{D\Delta}}{(\pi/2)^{1/4}} \approx 0.89 \frac{\sigma}{A} \sqrt{D\Delta}. \tag{28}$$

If $\sigma/A = 5\%$, $D = 4$, $\Delta = 1$, one has $\delta x_0 \approx 0.089$, i.e. about 11 times smaller than the bin width.

In the second case both charasteristics scales D and Δ are approximately of the same size

$$D \sim \Delta.$$

In this case the approximation Eq.(27) is not valid and one should compute explicitly the sum in the denominator of the Eq.(26).

One application of the formula (26) gives the answer to the very important question naturally posed by users of discrete detectors: if $D \sim \Delta$ holds, what is the best (minimal) limiting resolution $\delta x_0 = \sqrt{D(x_0)}$ in dependence of the input noise level?

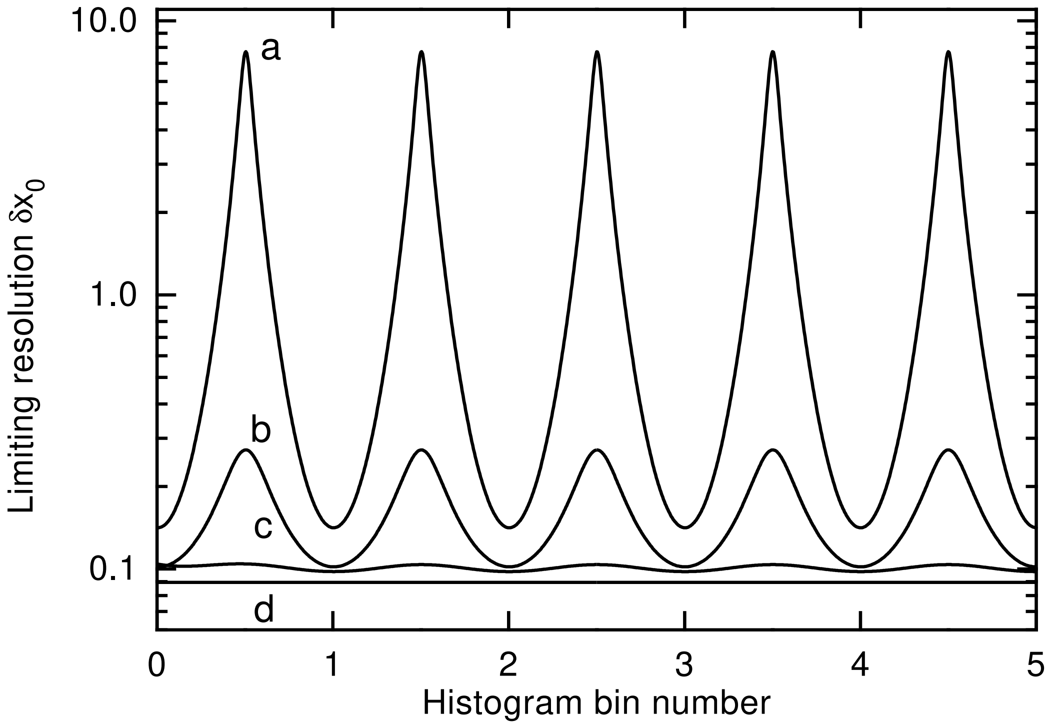

Fig. 2. Cramér-Rao limiting resolution δx_0 versus the signal $\psi(x - x_0)$ position x_0 on the X axis for 4 values of $D/\Delta = 1/4, 1/2, 1, \infty$ - the curves a, b, c and d respectively.

The answer can be obtained from Fig.2. It shows the series of curves presenting the limiting resolution δx_0 with the input noise level $\sigma/A = 10\%$ and bin number $n = 20$ for all graphics. The value of δx_0 is determined in correspondence with Eq.(26) by the formula

$$\delta x_0 \geq f(x_0; D, \Delta, n) \cdot \frac{\sigma}{A},$$

where the function $f(x_0; D, \Delta, n)$ is determined as:

$$f(x_0) = \frac{\sqrt{D\Delta}}{\left(\sum_{i=1}^{n}\left[\frac{\partial\psi(x_i - x_0)}{\partial x_0}\right]^2\right)^{1/2}}. \tag{29}$$

The curves a, b and c correspond to three values of the ratio $D/\Delta = 0.25, 0.5, 1$. The curve d corresponds to the asymptotic case $D/\Delta \gg 1$ when the approximation (27) is valid. Positions of the signal maxima x_0 are indicated on the axis X. Points $x_0 = 0, 1, 2, 3, 4, 5$ correspond to the borders of adjacent bins.

As can be seen from Fig.2, the value of the minimum error when $D = \Delta$ (the curve c) is almost coincident with the asymptotic case (the curve d). While the width D of the gaussian instrumental function decreases in respect to the bin-size Δ, the value of the minimum error begins to be strongly dependent on the position of x_0 on the axis X. The error reaches its minimum on a border of adjacent bins while its maximum occurs in the middle of a bin. If the peak width is increasing, the error dependence on the peak position on the X axis is getting smoother, reaching its asymptotic value almost in $D = \Delta$. A similar dependence was apparently observed in [3], where the momentum method was applied for estimating the position of the peak maximum. Now one can see that such a conclusion has the general character of and is valid for arbitrary methods of the peak maximum position estimation.

3.3 Double peak resolution method

As it was pointed out above, there is a probable case when the extracted cluster is created by two (or even more) signals. The NP algorithm is able to produce the indication of the peak number and some estimates of their parameters, which can be used as initial values for the parametric algorithm. However, if the probability of the overlapping of three or more signals is negligible, one can test the shape of the registered histogram by its correspondence to one or two signals without applying NP methods. Such a test can be constructed from the second and third central moments of the histogram (see [13]).

Thus we have to generalise the above methods and algorithms to multi-signal cases. The shape of a histogram produced by two superposing peaks can be described as

$$f(x; A_1, x_1^{(0)}, A_2, x_2^{(0)}) = A_1 \exp\left[-\frac{(x - x_1^{(0)})^2}{2\sigma^2}\right] + A_2 \exp\left[-\frac{(x - x_2^{(0)})^2}{2\sigma^2}\right], \tag{30}$$

This expression depends on four parameters. To find them we have to minimize a functional generalising the Eq.(22)

$$\mathcal{L}_4 = \sum_i \left(a_i - f(x_i; A_1, x_1^{(0)}, A_2, x_2^{(0)})\right)^2. \tag{31}$$

A direct generalisation of the PAM procedure on four-parameter functional (31) would lead to treating a 5D elliptic paraboloid. To avoid that we take into account that the partial derivatives of $\mathcal{L}_4$ with respect to A_1, A_2 are linear, so we can easily calculate both amplitudes by solving the system of two linear equations

$$\frac{\partial \mathcal{L}_4}{\partial A_{1;2}} = 0. \tag{32}$$

Then we can apply the above mentioned iterative PAM-procedure for minimizing $\mathcal{L}_4$ with respect to two remaining parameters $x_1^{(1)}, x_2^{(1)}$. The cardinal problem of this way of solution is the most accurate **choice of initial values of parameters**, since it determines the convergence and the speed of the iteration process. As pointed out above, we can use the estimations $x_1^{(0)}, x_2^{(0)}$ obtained by our NP algorithm [2].

After inserting them into the linear system (32) we can solve it to obtain $A_1^{(0)}, A_2^{(0)}$ and then, as it was noted above, apply the PAM-procedure for minimizing $\mathcal{L}_4$ to calculate the next iterative values of $x_1^{(1)}, x_2^{(1)}$. The whole procedure is repeated iteratively until the corrections become less than a prescribed value or the number of iterations attains its limits. Coming back to the problem of 2D-estimation of the SDD signal parameters we use the above solution to obtain k estimations $(\hat{A}_j, \hat{R}_j)$ for each $\hat{x}_j = R_j \varphi_j$. The sought for positions of the signal (R_0, x_0) can be calculated as the center of gravity of these k estimations:

$$R_0 = \frac{\sum_j \hat{A}_j \hat{R}_j}{\sum_j \hat{A}_j}, \quad x_0 = \frac{\sum_j \hat{A}_j \hat{x}_j}{\sum_j \hat{A}_j}. \tag{33}$$

To obtain its amplitude A we can minimize the function

$$L(A) = \sum_j \left(\hat{A}_j - N(\hat{x}_j, \hat{R}_j; A, x_0, R_0) \right)^2, \tag{34}$$

where the function $N(x, r; A, x_0, R_0)$ is taken from (1). This gives

$$\hat{A} = \frac{\sum_j \hat{A}_j \exp\left[-\frac{(\hat{x}_j - x_0)^2}{2{\sigma_x}^2} - \frac{(\hat{R}_j - R_0)^2}{2{\sigma_r}^2} \right]}{\sum_j \exp\left[-\frac{(\hat{x}_j - x_0)^2}{2{\sigma_x}^2} - \frac{(\hat{R}_j - R_0)^2}{2{\sigma_r}^2}) \right]}. \tag{35}$$

4 Simulation results

The method was tested using both the Monte-Carlo simulated and the real data. To simulate a histogram data set for testing we integrate sequentially the Gaussian sum (30) in each bin for various D. Then random normally distributed noise with σ_{noise} is added to each bin.

Two examples of resolving multiple peaks from the unimodal histogram by the NP method are presented in Fig.3-4. In Fig.3 two peaks lying apart only for one half of the bin width were simulated for σ_{noise}=0.1% of the mean amplitude value and signal half-width equal to 2 bin widths (Fig.3-a). In Fig.3-b it is shown how the NP-method resolves both peaks and the accuracy of PAM is presented as the distribution of deviations from the modelled and estimated positions of one of these peaks. The combined NP-PAM algorithm reconstructs pulse positions with accuracy of better than 0.01 of bin width.

[2] some easier ways are proposed in [13]

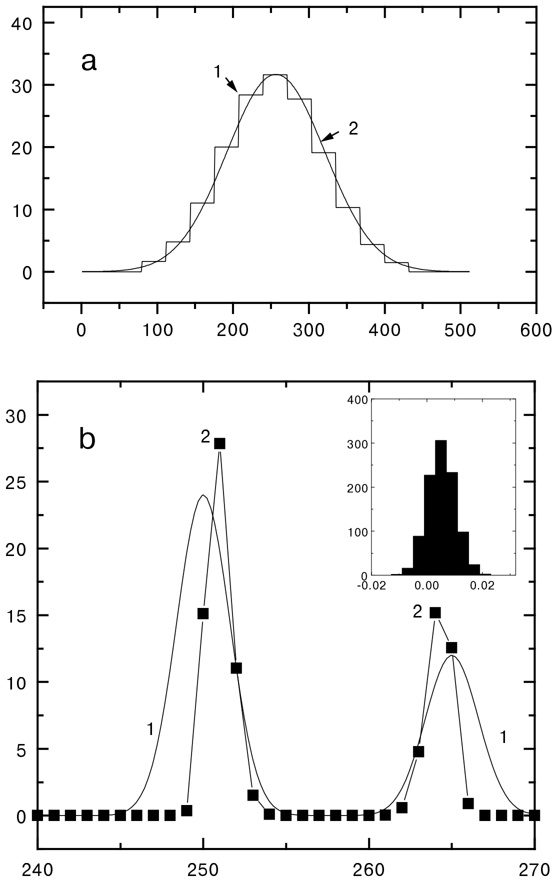

Fig. 3. The numerical test of two peak reconstruction from the 16 bins histogram, size of 1 bin equal 32; a) input data: 1 and 2 are marked correspondingly as the histogram and the instrumental function, $SNR = 60\ dB$; b) output data: the solid line 1 presents the original peaks, black squares 2 present peaks reconstructed by DCONV program [19]. The insert shows the distribution of deviations between simulated and estimated peak positions obtained by the parametric PAM algorithm with starting values provided by the NP algorithm.

An even more impressive example of three peak resolution from the unimodal histogram is given in Fig.4

The qualitative characteristics of corresponding deviations between the true data and the reconstructed ones by the DCONV program are given in Table 1.

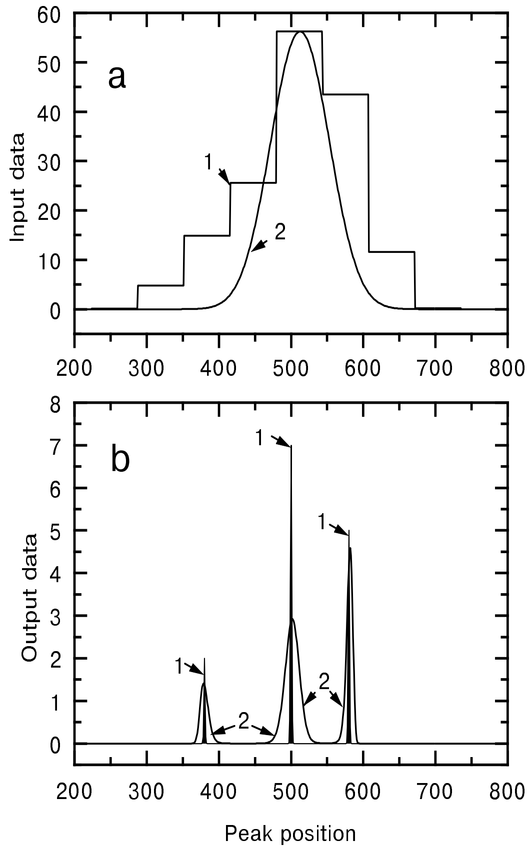

Fig. 4. Three peak reconstruction by the same program DCONV as in Fig.3: a) input data: histogram is marked by 1, instrumental function - by 2; b) output data: lines 1 present the original peaks, lines 2 present the reconstructed peaks.

Table 1.
The actual deviation of the line positions and the area under each of three peaks in the example presented at Fig.4

	Relative error of reconstruction (%)		
position	0.07	0.33	0.12
area under the peak	0	2.7	3.8
peak No.	1	2	3

The accuracy dependence of the COG and PAM procedures on the noise standard deviation is shown in Fig.5. The simulation was done for single symmetrical signals in order to have the best COG performance, but its accuracy is twice as bad as the PAM. The bottom curve presents the limiting accuracy determined by the Cramér-Rao formula (28). As one can see for σ_{noise} less than 10% the PAM procedure reaches almost its limiting accuracy.

Results of the comparative testing of the COG and PAM methods on double overlapping asymmetrical signals are presented in Fig.6. To provide a data set for testing a simple routine is written, which simulates histograms according to (20) for two overlapped "asymmetrical"

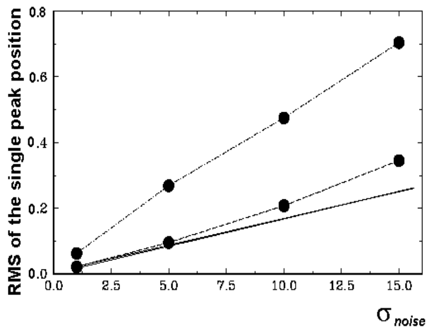

Fig. 5. RMS of the single signal position estimated by COG method (upper curve), by PAM-method (middle curve) and the Cramér-Rao lower bound (lower curve) as the function of the noise standard deviation. Both axes are measured in bin width

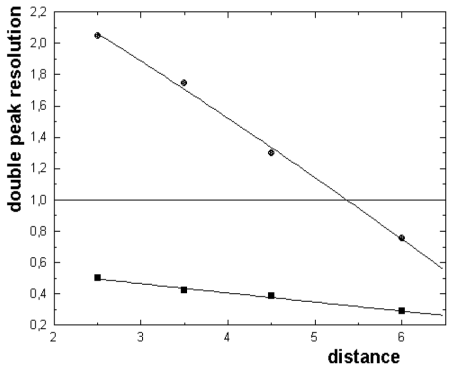

Fig. 6. Double pulse resolution versus distance between overlapped pulses (asymmetrical model). The upper curve presents results of the COG method, the lover curve shows results of PAM method. Axis units are in bin width. The line $RMS = 1$ separates the area of unacceptable results (greater than one bin width).

Gaussians with $\sigma_0 = 1.626 + 0.2097 \cdot 10^{-2} x_0$ and corresponding Δ. Contributions of every signal into each bin are superposed (summed). Then to each bin a random noise is added with the normal distribution with $\sigma = 10\%$ of the mean amplitude value of pulses. As it is seen in Fig.6, the COG double peak resolution is unsatisfactorily large exceeding the bin width on distances between peaks even greater than $3\sigma_0$ (see the part of the upper curve lying above the line RMS=1). However, the main defect of the COG estimation is its statistical bias that range is around 1.3-1.5 of bin width.

Besides the above shown accuracy dependence we also studied simulated data for the dependence of the PAM algorithm efficiency upon the gap between pulses and found that efficiency of the double pulse reconstruction is on the level of 93% for gaps greater than $1.5\sigma_0$ (where σ_0 is the Full Width of Half Maximum = FWHM). However it dramatically decreases to the level of 8% for gaps in the range $(0.7 - 1.5)\sigma_0$. In order to compare the quality of the PAM minimization procedure with the well-known MINUIT package [21] we consider calculations of the double pulse resolution by both programs as for simulated data as for the CERES Pb-Au'95 data. Our calculations show that the PAM algorithm in comparison with MINUIT has the same or even better accuracy for the double peak parameter reconstruction being 5-7 times faster in the range of distances from 7 time-bin up to $1.5\sigma_0$. Moreover, while PAM can process events with gaps between pulses equal even to 2 bin widths, the MINUIT method can not process any case when this distance is smaller or equal to 2.5. Fig.7 shows the distributions of the mean value of χ^2 as the function of the distance between pulses (normalised to σ) obtained by MINUIT and PAM for Pb-Au'95 data [13]. As one can see, the χ^2 - values of PAM are much lower than such values for the MINUIT method. This proves the better performance of the PAM fit.

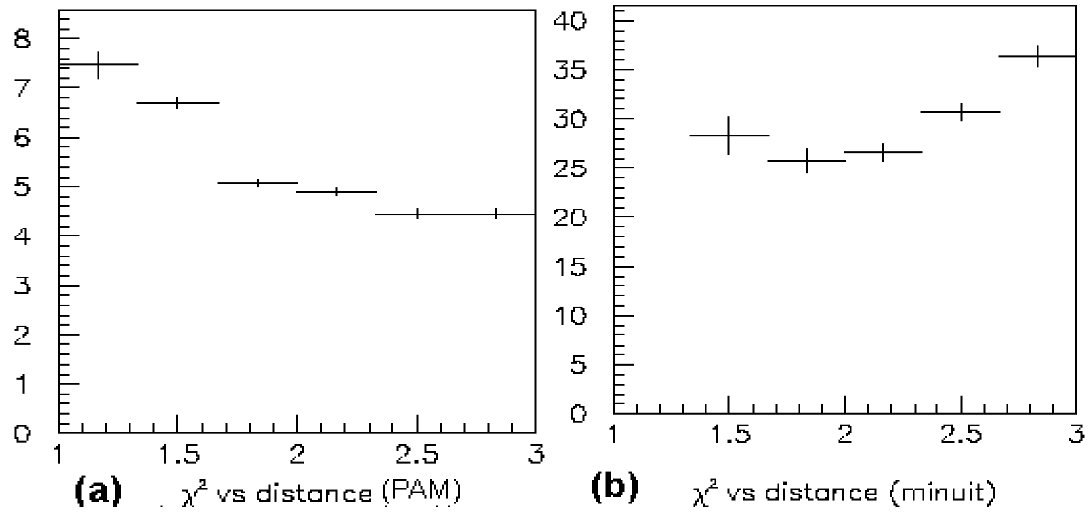

Fig. 7. χ^2 distribution for PAM (a) and MINUIT (b) versus distance between pulses (normalized to σ).

5 Conclusion

Various parametric and non-parametric approaches to the effective and precise determination of characteristics of signals registered by discrete detectors were studied in conditions of contemporary HEP experiments. The study was mainly focused on the quite delicate problem of the fast and most accurate resolution of overlapping signals.

It is proved that the recovery of particle coordinates by detectors with intrinsic resolution determined by the instrumental function and finite size of detector bins can be reduced to a non-parametric solution of the standard convolution integral equation with the kernel obtained by the convolution of the given instrumental function and the characteristic function of the histogram bin.

It is shown that a non-parametric method implemented by the modified DCONV routine of the RECOVERY program package is a powerful tool, which for known instrumental function allows resolution not only of doublets, but also triplets of signals being so close that their superposition was registered as a unimodal histogram. One should keep in mind, however, that these remarkable features have a natural resolution limit determined by the existing noise level. It follows from the close analogy between processing of noisy experimental data and the Shannon theorem on the highest possible transmission rate of information through a noisy channel. Detailed consideration of this analogy given above allows us to calculate a minimum attainable signal resolution for the given SNR (signal-to-noise-ratio).

Aiming for better accuracy and speed, the PAM (parabolic approximation method) is proposed. It is iterative and can use either NP or COG methods for providing the initial values of parameters. The comparative study of the PAM performance shows that it surpasses both the COG method in accuracy and the more sophisticated method used by the MINUIT package in speed and efficiency.

The Cramér-Rao inequality is applied to derive the lower-bound expression for the accuracy limit of the signal registered by a discrete detector with the given granularity and the noise level. It allows to appreciate the PAM performance, which appears surprisingly close to this limiting accuracy.

One of the PAM advantages is its possibility to be generalized for a non-gaussian shape of the signal. It is important since in reality signals can be symmetrical, but not gaussian, or slightly asymmetrical. For the latter case a signal model as the combination of two gaussians with different σ's is proposed. To be closer to real data the coefficients of this model have been fitted on the representative sample of the real Pb-Au data and, then used for simulating of the sample with signal doublets. The comparison of PAM and COG results obtained on this sample shows again the satisfactory accuracy and efficiency of the PAM and confirms that despite its popularity the COG methods should not be applied in any case where signal asymmetry or/and overlapping can be expected.

Acknowledgement

The authors are grateful to J.P.Wurm of Heidelberg University, Germany and Yu.A.Panebratsev of Joint Institute for Nuclear Research, Russia for useful discussions and valuable support in preparing this work. We thank E.Pantos and E.A.Hughes of CLRC, Daresbury Laboratory, UK, for the interest in this work and for their assistance with the preparation of the manuscript. The authors from JINR would like to acknowledge the support of Russian Foundation for Basic Research under Grant No. 970101027.

References

[1] J.M.F.Chamayou. *Averaging shifted histograms*, *Computer Physics Communications*, Vol.21, pp.145-161, 1980

[2] E.L.Kosarev, V.D.Peskov and E.R.Podolyak. *High resolution soft X-ray spectrum reconstruction by MWPC attenuation measurements*, Nucl.Instr. and Methods, Vol.208, pp.637-645, 1983

[3] W.R.Falk. *Data reduction from experimental histograms*, *Nucl.Instr. and Meth.*, Vol.220, pp.473-478, 1984

[4] E.Gatti, P.Rehak, M.Sampietro. *Double particle resolution in semiconductor drift detectors*, *Nucl.Instr. and Meth.,* Vol.A274, pp.469-476, 1989

[5] W.Chen et al. *Large Area Cylindrical Silicon Drift Detector*, *IEEE Trans. Nucl. Sci.,* Vol.NS-39, No 4, pp.619-628, 1992

[6] G.Stefanek. *Resolution of cylindrical Si-drift detector studied by using a computational model*, ALICE/93-24, Internal Note/SIL 31 May 1993, CERN, 1993 Vol.A338, pp.498-505, 1994

[7] E.Conti et al. *Performance of a liquid xenon time projection chamber for low energy γ-rays detection*, *Nucl.Instr. and Meth.,* Vol.A356, pp.286-296, 1995

[8] V.Ladygin, P.K.Manyakov, N.M.Piskunov. *Time-of-flight trigger with digital selection of events*, *Nucl.Instr. and Meth.,* Vol. A357, pp.386-390, 1995

[9] V.Gratchev et al. *Double track resolution of cathod strip chamber*, *Nucl.Instr. and Meth.,* Vol.A365, pp.576-584, 1995

[10] M.Faisal et al. *Implementation of a modified Richardson-Lucy method for image restoration on a massively parallel computer to compensate for space-variant point spread of a charge-coupled device camera*, *J.Opt.Soc.Am.A*, Vol.12, No.12, pp.2593-2603, 1995

[11] R.Bellwied et al. *Development of large linear silicon drift detectors for the STAR experiment at RHIC*, *Nucl. Instr. and Meth.,* Vol.A377, pp.387-392, 1996

[12] G.Ososkov, A.Shitov,*Gaussian Wavelet Features and Their Applications for Analysis of Discretized Signals*, Proc of MTCP98, Comp. Phys. Coomm., (1999) (in press, see also Commun. of JINR P11-97-347, Dubna, 1997).

[13] G.Agakishiev, E. Kolganova, G.Ososkov et al. *Effective Pulse Resolution Algorithms for Detectors with Gaussian-like Signal Shape*, JINR Comm. E10-97-105, Dubna, (1997), 16p.

[14] C.E.Shannon, *Communication in the presence of noise*, Proc.of the IRE, Vol.37, No.1, pp.10-21, 1949

[15] N.Chernov, E. Kolganova, G.Ososkov, *Robust Methods for RICH Ring Recognition and Particle Identification*, in: Proc. RICH'98, Nucl. Instr. and Meth. (in press, see also Commun. of JINR E10-95-468, 1995).

[16] U.Fashingbauer, G.Agakishiev, R.Baur, et al. *A doublet of 3 inch cylindrical silicon drift detectors in the CERES experiment*, *Nucl. Instr. and Meth.,* Vol.A377, No.2-3, pp.362-366, 1996

[17] A.N.Tikhonov, V.Ya.Arsenin. *Solution of ill-posed problems*, Wiley, 1977

[18] E.L.Kosarev. *Shannon's superresolution limit for signal recovery*, *Inverse Problems*, Vol.6, No.1, pp.55-76, 1990

[19] V.I.Gelfgat, E.L.Kosarev, E.R.Podolyak. *Programs for signal recovery from noisy data using the maximum likelihood principle. Part 1: General description*, *Computer Physics Communications*, Vol.74, No.3, pp.335-348, 1993, Part 2: Program implementation. ibid, pp.349-357, 1993

[20] C.R.Rao. *Linear statistical inference and its applications.* Wiley, 1965

[21] F.James. *MINUIT - Function Minimization and Error Analysis*, CERN Program Library entry D506

Figure captions

Fig.1 CERES SDD image of an Au-Pb event.

Fig.2 Cramér-Rao limiting resolution δx_0 versus the signal $\psi(x - x_0)$ position x_0 on the X axis for 4 values of $D/\Delta = 1/4, 1/2, 1, \infty$ - the curves a, b, c and d respectively.

Fig.3 The numerical test of two peak reconstruction from the 16 bins histogram, size of 1 bin equal 32; a) input data: 1 and 2 are marked correspondingly as the histogram and the instrumental function, $SNR = 60\ dB$; b) output data: the solid line 1 presents the original peaks, black squares 2 present peaks reconstructed by DCONV program [19]. The insert shows the distribution of deviations between simulated and estimated peak positions obtained by the parametric PAM algorithm with starting values provided by the NP algorithm.

Fig.4 Three peak reconstruction by the same program DCONV as in Fig.3: a) input data: histogram is marked by 1, instrumental function - by 2; b) output data: lines 1 present the original peaks, lines 2 present the reconstructed peaks.

Fig.5 RMS of the single signal position estimated by COG method (upper curve), by PAM-method (middle curve) and the Cramér-Rao lower bound (lower curve) as the function of the noise standard deviation. Both axes are measured in bin width.

Fig.6 Double pulse resolution versus distance between overlapped pulses (asymmetrical model). The upper curve presents results of the COG method, the lover curve shows results of PAM method. Axis units are in bin width. The line $RMS = 1$ separates the area of unacceptable results (greater than one bin width).

Fig.7 χ^2 distribution for PAM (a) and MINUIT (b) versus distance between pulses (normalized to σ).

Table 1.
The actual deviation of the line positions and the area under each of three peaks in the example presented at Fig.4

	Relative error of reconstruction (%)		
position	0.07	0.33	0.12
area under the peak	0	2.7	3.8
peak No.	1	2	3