ELSEVIER

Nuclear Instruments and Methods in Physics Research A 440 (2000) 431–437

IN PHYSICS RESEARCH Section A

www.elsevier.nl/locate/nima

Constrained minimization in the C ++ environment

S.N. Dymov, V.S. Kurbatov, I.N. Silin, S.V. Yaschenko*

Laboratory of Nuclear Problems, Joint Institute for Nuclear Research, 141980 Dubna, Russia

Received 11 December 1998; received in revised form 24 June 1999 accepted 6 July 1999

Abstract

On the basis of the ideas, proposed by one of the authors (I.N. Silin), a suitable software has been developed for constrained data fitting. Constraints may be of arbitrary type; i.e. equalities and inequalities. The simplest possible way has been used. The widely known program FUMILI was re-written in the C ++ language. Constraints in the form of inequalities $\phi(\boldsymbol{\theta}) \geq a$ were taken into account by changing them into equalities $\phi(\boldsymbol{\theta}) = t$ and simple inequalities of type $t \geq a$. The equalities were taken into account by means of quadratic penalty functions. The suitable software was tested on the model data for the ANKE setup (COSY accelerator, Forschungszentrum Jülich, Germany). © 2000 Elsevier Science B.V. All rights reserved.

Keywords: FUMILI; ANKE; Codes

1. Introduction

In the present paper we describe two realizations (in C ++ language) of constrained minimization for χ^2-like functionals. One of them is the algorithm of the FUMILI code, which was available for users as a part of CERN library [1]. The description of this algorithm was published in Russian [2] at the end of the 1960s. Due to the fact that the access to this publication is not easy for an English reader, we give a short description of the FUMILI algorithm. This algorithm is now coded in the C ++ language.

The second part is the realization of the idea proposed by one of the authors (I.N. Silin) for solving the constrained minimization problem in a general case, where constraints are of arbitrary type (arbitrary equalities and inequalities) [3]. Technically, here constraints are taken into account by the method of penalty functions (though there are other ways of doing this [3]). The algorithm described below was tested on the model data for the calibration process $pp \rightarrow d\pi^+$ under the conditions of the ANKE setup [4].

2. Algorithm of FUMILI

For simplicity, let us assume that the function to be minimized has the form[1]

$$\chi^2 = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{f_j(\boldsymbol{x}_j, \boldsymbol{\theta}) - F_j}{\sigma_j}\right)^2 \qquad (1)$$

* Corresponding author.

[1] What follows can be easily generalized to the case where the covariance matrix of the data F_j has non-diagonal terms.

PII: S0168-9002(99)00758-5

where $f_j(\boldsymbol{x}_j, \boldsymbol{\theta})$ are the measured functions at the points $\boldsymbol{x}_j$, F_j are the measured values, σ_j are their errors, and $\boldsymbol{\theta}$ are parameters to be estimated.

The minimum condition is

$$\frac{\partial \chi^2}{\partial \theta_i} = \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \cdot \frac{\partial f_j}{\partial \theta_i} [f_j(\boldsymbol{x}_j, \boldsymbol{\theta}) - F_j] = 0, \quad i = 1, \dots, m \tag{2}$$

where m is the number of parameters.

Expanding the left hand side of Eq. (2) in parameter increments and retaining only linear terms we get

$$\left(\frac{\partial \chi^2}{\partial \theta_i}\right)_{\theta=\theta^0} + \sum_k \left(\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_k}\right)_{\theta=\theta^0} \cdot (\theta_k - \theta_k^0) = 0.$$

Here $\boldsymbol{\theta}^0$ is some initial value of parameters. In a general case

$$\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_k} = \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \cdot \frac{\partial f_j}{\partial \theta_i} \cdot \frac{\partial f_j}{\partial \theta_k} + \sum_{j=1}^{n} \frac{(f_j - F_j)}{\sigma_j^2} \cdot \frac{\partial^2 f_j}{\partial \theta_i \partial \theta_k}. \tag{3}$$

In the FUMILI algorithm an approximate expression (3) for $\partial^2\chi^2/\partial\theta_i\partial\theta_k$ is used in which the last term is discarded (it is often done, not always wittingly, and sometimes causes trouble), i.e.:

$$\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_k} \cong Z_{ik} = \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \cdot \frac{\partial f_j}{\partial \theta_i} \cdot \frac{\partial f_j}{\partial \theta_k}.$$

As a result the equations for parameter increments have the following form:

$$\left(\frac{\partial \chi^2}{\partial \theta_i}\right)_{\theta=\theta^0} + \sum_k Z_{ik} \cdot (\theta_k - \theta_k^0) = 0, \quad i = 1, \dots, m.$$

A remarkable feature of the algorithm is the technique used for step restriction. For the current approximation of the optimal parameters $\boldsymbol{\theta}^0$ a parallelepiped P_0 is built with a centre at $\boldsymbol{\theta}^0$ and axes parallel to the coordinate axes θ_i. The lengths of the parallelepiped sides along the i-th axis are $2 \cdot b_i$, where b_i have such values that the functions $f_j(\boldsymbol{\theta})$ are quasi-linear all over the parallelepiped. If the step $\Delta\boldsymbol{\theta}$ gives a new point $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \Delta\boldsymbol{\theta}$ outside P_0, the crossing $\boldsymbol{\theta}^1$ of the vector $\Delta\boldsymbol{\theta}$ with the surface of P_0 is found and taken as a new value for the parameters. After selection of a new value for the parameters, it is checked whether the function reduction is big enough compared with that expected in the quadratic approximation. If it is not, the step reduction is performed. Some parallelepiped lengths can be increased too.

In addition, FUMILI takes into account simple linear inequalities such as

$$\theta_i^{\min} \le \theta_i \le \theta_i^{\max}. \tag{4}$$

They form a parallelepiped P (P_0 may be deformed by P). If the value of the parameter coincides with its restriction (the approximation lies on the surface of P) and the gradient component is such that χ^2 is not going to increase beyond P, the corresponding parameter is fixed.

Then the step is calculated for all non-fixed parameters and if some parameters, lying on the surface of P, go beyond P, one of them is temporarily fixed too (the parameter, for which the ratio $|\Delta\theta_i|/\sqrt{(Z^{-1})_{ii}}$ is maximal), and so on.

The criterion to be fulfilled for the end of the iteration process is that all parameters are fixed due to only the gradient component signs and step increments for non-fixed parameters

$$|\Delta\theta_i| < \varepsilon \cdot \sqrt{(Z^{-1})_{ii}}$$

where ε is a small number ~ 0.01. As the number of fixation combinations is finite, the number of steps will also be finite, at least in the convex quadratic case.

Very similar step formulae are used in FUMILI for the negative logarithm of the likelihood function with the same idea of linearizing the functional argument.

3. Minimization of χ^2 functionals with arbitrary constraints

3.1. Formulation of the problem

Again, let us assume that the function to be minimized has the same form (1), but in addition to simple linear constraints (4) there are two more types of constraints, i.e. non-linear inequalities and equalities:

$$a_r \le \phi_r(\boldsymbol{\theta}) \le b_r, \quad r = 1, \dots, m_d \tag{5}$$

$$\psi_s(\boldsymbol{\theta}) = c_s, \quad s = 1, \dots, m_e. \tag{6}$$

Here $\phi_r(\boldsymbol{\theta}), \psi_s(\boldsymbol{\theta})$ are the regular functions of the parameter $\boldsymbol{\theta}$; a_r, b_r, m_d are the low and upper boundaries of the inequalities and their number; c_s, m_e are any constant and number of equations. The regularity is taken to mean continuous second-order derivatives. The problem of taking into account the constraints in the form of the equalities of type (6) was solved before [5–7]. As for the constraints in the form of inequalities (5), the authors did not know a simple solution until one of them (I.N. Silin) proposed a method for taking them into account [3]. According to Ref. [3], any constraint of the form $a_r \le \phi_r(\boldsymbol{\theta}) \le b_r$ can be replaced by a simple inequality and equality, namely

$$a_r \le t_r \le b_r \tag{7}$$

$$\phi_r(\boldsymbol{\theta}) = t_r. \tag{8}$$

Here t_r is an additional variable constrained by two boundaries a_r, b_r, (8) is a constraint in the form of the equation. You can see that constraints (7) have the same form and structure as those of Eq. (4), so we can combine them and introduce just one type of simple constraint:

$$\theta_i^{\min} \le \theta_i \le \theta_i^{\max}$$

where index i changes in a wider range: $i = 1, \dots, m, \dots, m + m_d$ and for $i > m$

$$\theta_i = t_{i-m}, \quad \theta_i^{\min} = a_{i-m}, \quad \theta_i^{\max} = b_{i-m}.$$

Then the problem of constrained minimization in a general case can be reformulated as follows. We find a minimum of function (1) under the constraints:

$$\theta_i^{\min} \le \theta_i \le \theta_i^{\max}, \quad i = 1, \dots, m, \dots, m + m_d \tag{9}$$

$$\xi_u(\boldsymbol{\theta}) = d_u, \quad u = 1, \dots, m_d, \dots, m_d + m_e \tag{10}$$

where for $1 \le u \le m_d$, $\xi_u = \phi_u$, $d_u = t_u$ and for $m_d < u \le m_d + m_e$, $\xi_u = \psi_{u-m_d}$, $d_u = c_{u-m_d}$.

After such reformulation the number of the parameters to be fitted and of the constraints in the form of simple inequalities of type (9) becomes $m + m_d$; and the number of constraints in the form of equations of type (10) becomes $m_d + m_e$.

When non-simple constraints are only equations they can be taken into account by using either the method proposed in Ref. [7] or the penalty function method. Here we use the latter. In the penalty function method a minimum of function (11) is searched for as $T \to \infty$.

$$\Phi = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{f_j(\boldsymbol{x}_j, \boldsymbol{\theta}) - F_j}{\sigma_i}\right)^2 + \frac{1}{2}T\left(\sum_{r=1}^{m_d}\frac{(\phi_r - \theta_{r+m})^2}{\sigma_r^2} + \sum_{s=1}^{m_e}\frac{(\psi_s - c_s)^2}{\sigma_s^2}\right). \tag{11}$$

Here T is the penalty factor (normally it is a sufficiently big number), and σ_r, σ_s are the formally calculated errors of constraints.

3.2. Iteration scheme

Let us rewrite Eq. (11) in the form

$$\Phi = \Phi_1 + \frac{1}{2}\sum_{r=1}^{m_d}\frac{(\phi_r - \theta_{r+m})^2}{w_r}$$

where

$$\Phi_1 = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{f_j(\boldsymbol{x}_j, \boldsymbol{\theta}) - F_j}{\sigma_j}\right)^2 + \frac{1}{2}T\sum_{s=1}^{m_e}\frac{(\psi_s - c_s)^2}{\sigma_s^2}$$

and $w_r = \sigma_r^2/T$. With a chosen T the minimum condition is

$$\frac{\partial\Phi}{\partial\theta_k} = \frac{\partial\Phi_1}{\partial\theta_k} + \sum_{r=1}^{m_d}\frac{\partial\phi_r}{\partial\theta_k}\cdot\frac{(\phi_r - \theta_{r+m})}{w_r} = 0, \quad (k = 1, \dots, m) \tag{12}$$

$$\frac{\partial\Phi}{\partial\theta_{r+m}} = -\frac{(\phi_r - \theta_{r+m})}{w_r} = 0, \quad (r = 1, \dots, m_d). \tag{13}$$

In both Eqs. (12) and (13) derivatives are taken only for those parameters which are not fixed, i.e. $k \ne i_{\mathrm{f}}, r + m \ne i_{\mathrm{f}}$, where i_{f} is the index of a fixed parameters. The functions on the left-hand sides of Eqs. (12) and (13) depend on the $m + m_d$ parameters. Near the minimum we can expand the left-hand sides of the equations in parameter increments retaining only linear terms. For Eq. (12) we have

$$\left[\frac{\partial\Phi_1}{\partial\theta_k} + \sum_{r=1}^{m_d}\frac{\partial\phi_r}{\partial\theta_k}\cdot\frac{(\phi_r - \theta_{r+m})}{w_r}\right] + \sum_{l=1}^{m}\left[\frac{\partial^2\Phi_1}{\partial\theta_k\partial\theta_1} + \sum_{r=1}^{m_d}\frac{\partial\phi_r}{\partial\theta_k}\cdot\frac{\partial\phi_r}{\partial\theta_1}\cdot\frac{1}{w_r}\right]\cdot\delta\theta_1 - \sum_{r=1}^{m_d}\frac{\partial\phi_r}{\partial\theta_k}\cdot\frac{\delta\theta_{r+m}}{w_r} = 0. \tag{14}$$

We wrote Eq. (14) in the approximation of the functional argument linearization method [8], in which the derivatives $\partial^2\phi/\partial\theta_k\partial\theta_l$ are discarded. All values of functions and derivatives are taken for the current values of the parameters. Let us also note that index l ($l \neq i_f$) in the second term runs over indices of non-fixed parameters. By analogy, for Eq. (13).

$$[\phi_r - \theta_{r+m}] + \sum_{l=1}^{m} \frac{\partial\phi_r}{\partial\theta_l}\cdot\delta\theta_l - \delta\theta_{r+m} = 0. \qquad (15)$$

From Eq. (15), for the non-fixed parameter θ_{r+m} ($r = 1,\dots,m_d$) we have

$$\delta\theta_{r+m} = [\phi_r - \theta_{r+m}] + \sum_{l=1}^{m} \frac{\partial\phi_r}{\partial\theta_l}\delta\theta_l. \qquad (16)$$

Substituting Eq. (16) into Eq. (14) and after some algebra we obtain

$$G_k + \sum_{l=1}^{m} Z_{kl}\cdot\delta\theta_l = 0 \qquad (17)$$

where

$$G_k = \frac{\partial\Phi_1}{\partial\theta_k} + \sum_{r=1}^{m_d}\left[\frac{\partial\phi_r}{\partial\theta_k}\cdot\frac{(\phi_r - \theta_{r+m})}{w_r}\right]$$

$$Z_{kl} = \frac{\partial^2\Phi_1}{\partial\theta_k\partial\theta_l} + \sum_{r=1}^{m_d}\frac{\partial\phi_r}{\partial\theta_k}\cdot\frac{\partial\phi_r}{\partial\theta_l}\cdot\frac{1}{w_r}.$$

A remarkable feature of the last expressions is that the index l runs only over non-fixed parameters $l = 1,\dots,m$, and the index r runs only over those inequalities for which additional parameters t_r (7) are fixed!

Finally, the solution of Eq. (17) is

$$\delta\boldsymbol{\theta} = -(Z^{-1}\cdot G).$$

The increments of the additional parameters $\delta t_r = \delta\theta_{r+m}$ are calculated according to formula (16).

The advantage of this iteration scheme is that the matrix inversion only of order $m\times m$ is done irrespective of the number of constraints.

The scheme can be improved in order to avoid the influence of non-linear effects of permanently valid inequalities. To do this θ_{r+m} which is neither on low nor upper boundary is substituted by the quantity $\psi_r(\theta)$ (or its boundary value if the former lies outside the boundary) after every iteration step.

4. Test

Both realizations described above are coded in C + + and tested on the model data for the calibration reaction pp → dπ^+ under the conditions of the ANKE setup [4]. According to the plans, ANKE will consist of three detectors: a side detector, forward and backward ones. At the moment the side detector is fully assembled, only a scintillation hodoscope is ready for the forward detector. The side detector consists of two scintillation hodoscopes (START, STOP), and two proportional chambers each consisting of three sensitive planes. It allows one to reconstruct all the kinematic parameters of the ejectiles passing through the side detector. The scintillation hodoscope incorporated in the forward detector is capable of measuring the coordinates of the particle and its time of flight.

The first data were obtained in May and July 1998, the accuracies being studied. As the main calibration process required for the analysis of detector performance is the reaction pp → dπ^+, we took this process for the tests. A number of events was simulated for the beam kinetic energy $T_{\text{beam}} = 425$ MeV with the π^+ meson passing through the side detector and the deuterons passing through the scintillation hodoscope of the forward detector. Simulation was done by the GEANT code with all physical processes switched on except the decay of π^+ mesons. In the case where the kinematic parameters of the beam proton and secondary π^+ are known, there is one constraint having the form of an equality, namely, the missing mass of the process should be equal to the mass of the deuteron:

$$(E_{\text{beam}} + M_{\text{p}} - E_{\pi^+})^2 - (\boldsymbol{p}_{\text{beam}} - \boldsymbol{p}_{\pi^+})^2 = M_{\text{d}}^2 \qquad (18)$$

where $E_{\text{beam}}, E_{\pi+}$ are the energies of the beam proton and the secondary π^+ meson, $\boldsymbol{p}_{\text{beam}}, \boldsymbol{p}_{\pi^+}$ are their 3-momenta, $M_{\text{p}}, M_{\text{d}}$ are the masses of the proton and the deuteron, respectively.

As noted above, for the deuterons detected by the forward hodoscope, their coordinates and times of flight (t_{d}) will be measured hopefully with the accuracies permitting 4c fit (using all 4 conservation laws). As at the moment not all accuracies are known, we assume that their coordinates and times of flight lie between some boundaries and put forward requirements having the form of the following three inequalities:

$$y_{\min} \le y_d \le y_{\max}, \quad z_{\min} \le z_d \le z_{\max},$$
$$t_{\min} \le t_{\mathrm{d}} \le t_{\max}. \tag{19}$$

The first two requirements come from the geometrical dimensions of the scintillation hodoscope, whereas the last one does from the simulation data. Three functions y_d, z_d, t_d were expressed as the ones (in the form of polynomials to the third-order inclusive) of two angles of the pion in the laboratory system of coordinates.

The total number of fitted parameters was six, the first three are angles θ_{xz}, θ_{yz} of the pion relative to the beam proton and the pion momentum in the laboratory system. The last three parameters were additional parameters t_r, corresponding to three inequalities (19). Initial pion angles were always 0, and initial momenta were calculated as a function of these angles. The coordinates of the pion detected in the side detector were expressed as functions of three pion variables, i.e., two angles and momentum. The total number of events was ~ 3000, the maximum number of iterations was 40.

Two fits corresponding to two different realizations, described above, were performed. In the first fit the constraint of the form of non-linear equation (18) was disabled, while in the second one it was enabled. In Fig. 1 the accuracies for both realizations are shown. Figs. 1(a)–(c) are for the first fit, Figs. 1(d)–(f) are for the second one. It is necessary to stress a drastic improvement of accuracy in $\Delta p/p$ in the second case, which is the result of additional constraint.

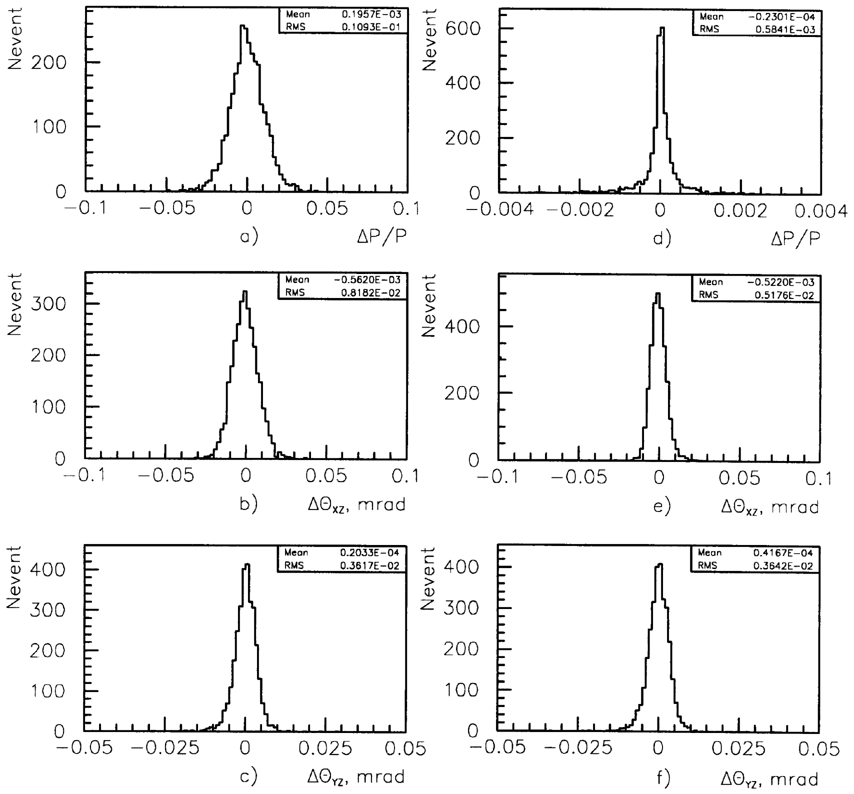

Fig. 1. Accuracies of determining the particle kinematic parameter.

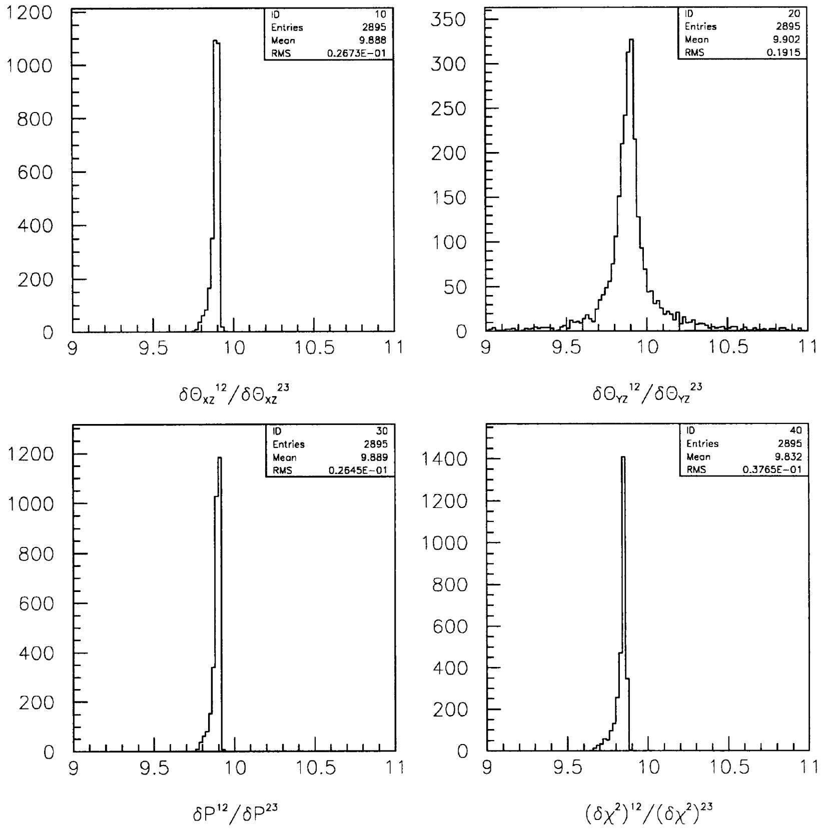

Fig. 2. Illustration of the Richardson approximation.

Each event was fitted to three values of the penalty factor T. The initial value of this factor was selected by using the formula

$$T = 100 \cdot \frac{n_{\text{exp}}}{n_{\text{con}}}$$

where n_{exp} is the number of experimental points, and n_{con} is the number of constraints. In our case $n_{\text{exp}} = 6$, $n_{\text{con}} = 4$.

Each successive value of T was ten times larger than the previous one. According to Ref. [3], in this case, according to Richardson, we should have the convergence, i.e. the parameter improvements $\Delta_{32} = p_3 - p_2$ should be 10 times smaller than $\Delta_{21} = p_2 - p_1$. Here p_1, p_2 and p_3 mean the values of the fitted parameter for T equal to T_1, T_2 and T_3, respectively.

In Fig. 2 the ratios $\Delta\theta_{xz}^{32}/\Delta\theta_{xz}^{21}$, $\Delta\theta_{yz}^{32}/\Delta\theta_{yz}^{21}$, $\Delta p^{32}/\Delta p^{21}$, $\Delta(\chi^2)^{32}/\Delta(\chi^2)^{21}$ are shown. It is seen that they are close to 10, and this indicates that the statements made in Ref. [3] are correct.

5. Conclusion

Two codes are developed for the minimization of χ^2-like functionals in the C + + language. One of them is realization of the FUMILI code with constraints of the form of simple boundaries. The second one is the minimization with constraints of any type. With FUMILI as a starting point, the C + + code is developed and tested on model data. The results of the test show a high performance of the algorithms developed. In conclusion, the authors express their gratitude to their colleagues from the

ANKE collaboration for providing the necessary details and to L. Pashkevich for her help in preparing an English version of the paper.

References

[1] I.N. Silin, CERN Program Library D 510, FUMILI, 1983.

[2] I.N. Silin, Appendix III, Statisticheski metodi v eksperimentalnoi fizike, Atomizdat 1976 (translated into Russian from W.T. Eadie et al., Statistical Methods in Experimental Physics, CERN, Geneva, 1971, North-Holland, Amsterdam).

[3] I.N. Silin, Resolute progress in a constrained minimization problem, JINR Rapid Communications, No 3 [89]-98, pp. 25–30.

[4] Forschungszentrum Jülich, Germany, Annual Report 1996 (all the details about ANKE setup and related bibliography can be found in this report).

[5] J.P. Berge, F.T. Solmitz, H.D. Taft, Rev. Sci. Instr. 32 (1961) 538.

[6] R. Bock, CERN 60-30 (1960).

[7] V.S. Kurbatov, I.N. Silin, Nucl. Instr. and Meth. A 345 (1994) 346.

[8] S.N. Sokolov, I.N. Silin, Preprint JINR D-810, Dubna, 1961.