объединенный
институт
ядерных
исследований
дубна

E1-93-253

V.S.Kurbatov, I.N.Silin

NEW METHOD FOR MINIMIZING REGULAR FUNCTIONS WITH CONSTRAINTS ON PARAMETER REGION

Submitted to «Nuclear Instruments and Methods A»

1993

1. Introduction

Approximately thirty years ago a linearization method for minimizing χ^2 - like functionals was proposed [1], subroutine FUMILI was developed by one of the authors (I.N. Silin) [2] and became available for users. Few years later I.N. Silin implemented in FUMILI the simplest case of constrained fit for constraints of type $a \leq X \leq b$. Unfortunately at CERN during adapting FUMILI to new versions of FORTRAN this option was in fact lost : by default at the very beginning each parameter gets a equal to the smallest number;and b. to largest one. Such a setting is done in BLOCK DATA statement. But by mistake this setting was moved from BLOCK DATA to FUMILI itself, so when a user sets his bounds before call to FUMILI they are automatically erased at the entry to FUMILI.

As a rule, FUMILI efficiently mininiizes χ^2 - like functions (including the logarithm of general type likelihood functions) which are in fact functionals F_U on the discrete set $Y_U(X)$ of the functional argument $y(U, X)$ with the values defined by a finite number of parameters X, i.e. $F_U(y(U,X)) = \sum_i f(y_i(u_i, X))$. But sometimes it gets into trouble and the main reason is in the following : in FUMILI as the second derivatives of minimized function their approximate values are used with the neglection of members containing second derivatives of the functional argument. For χ^2 - like functionals such matrix of second derivatives is nonnegatively defined. The problem appears when approximate second derivatives matrix has eigenvalues close or equal to zero. In this case FUMILI not only cannot find the minimum but simply descend to the lower values of the function. Nevertheless under full unconditionness of the matrix the parameters responsible for this are being fixed. They may be fixed by the user too.

Firstly, the degeneration may happen when the user tries to determine too many parameters.

Secondly (it is the worst case), degeneration of the matrix happens to be when the first derivative of the functional argument over fitted parameters or any of their linear combination becomes equal to zero on the whole set U. Such degeneration may happen, for example, when one tries to take into account the linear bounds by change of variables and completely destroys convergence.

In addition to this FUMILI cannot minimize functions of arbitrary structure. During many years I.N. Silin worked on a new algorithm with the aim to overcome these restrictions. The work on such an algorithm was intensified when another author (V.S.Kurbatov) built a practically acceptable algorithm for reducing the order of the problem by taking into account nonlinear constraints [3] and [4]. The more constraints exists, the more stable solution search must be, if we do it accurately.

The idea is very simple. Let us assume that we have a quadratic function

$$F = F(X_0) + \sum G_i \cdot (X_i - X_i^0) + \frac{1}{2} \sum_{i,j} (X_i - X_i^0) \cdot Z_{i,j} \cdot (X_j - X_j^0)$$
$$= F_0 + \mathbf{G} \cdot \Delta\mathbf{X} + \frac{1}{2} \Delta\mathbf{X}^T \cdot \mathbf{Z} \cdot \Delta\mathbf{X} \qquad (1)$$

and the constraints

$$f_\lambda(\mathbf{X}) \; = \; 0; \; \lambda = 1 \div nc \qquad (2)$$

(superscript T means transposition).

If constraints are regular functions of parameters we may linearize them in the vicinity of $\mathbf{X}_0$.

$$\mathbf{f} = \mathbf{f}(\mathbf{X}_0) + \mathbf{DF} \cdot \Delta\mathbf{X}. \qquad (3)$$

Here $\mathbf{DF}$ is the rectangular matrix $(nc \cdot np)$ of first derivatives of constraint functions over parameters $\mathbf{X}$ at $\mathbf{X} = \mathbf{X}_0$. We can subdivide vector $\Delta\mathbf{X}^T = (\Delta\mathbf{X1}^T, \Delta\mathbf{X2}^T)$ where $\Delta\mathbf{X1}^T$ has $(np - nc)$ components and $\Delta\mathbf{X2}^T$ has nc components. The main trick is how to do subdivision of vector $\mathbf{X}$ into two subvectors. We shall talk about it later.

Using the previous equation we can express

$$\Delta\mathbf{X2} = \mathbf{R} + \mathbf{S} \cdot \Delta\mathbf{X1} \qquad (4)$$

After substitution (4) into (1) we will obtain another quadratic approximation of the minimized function

$$F = F_0' + \mathbf{G}' \cdot \Delta\mathbf{X1} + \frac{1}{2} \Delta\mathbf{X1}^T \cdot \mathbf{Z}' \cdot \Delta\mathbf{X1}$$
$$F_0' = F(X_0) + \sum_{i=1}^{nc} R_i \cdot [G_{px2(i)} + \frac{1}{2} \sum_{j=1}^{nc} Z_{px2(i),px2(j)} \cdot R_j]$$
$$G'_i = G_{px1(i)} + \sum_{j=1}^{nc} [G_{px2(j)} \cdot S_{j,i} + R_j \cdot [Z_{px1(i),px2(j)} + \sum_{k=1}^{nc} S_{k,i} \cdot Z_{px2(k),px2(j)}]]$$
$$Z'_{i,j} = Z_{px1(i),px1(j)} +$$
$$\sum_{k=1}^{nc} [S_{k,i} \cdot Z_{px2(k),px1(j)} + S_{k,j} \cdot Z_{px2(k),px1(i)} + S_{k,i} \cdot \sum_{l=1}^{nc} Z_{px2(k),px2(l)} \cdot S_{l,j}] \, . \qquad (5)$$

Here we use the following notation : $px1(i)$ - index function, meaning the parameter number of the i-th component of vector $\mathbf{X1}$, $px2(i)$ - index function, meaning the parameter number of the i-th component of vector $\mathbf{X2}$. Using such a technique we can take into account the constraints of general type : both the inequalities and equalities[4]

At the end a new code was created and called FUMIVI : FUnction MInimization by Vallies Investigation. The main features of the new code :

- Minimization of regular functions of arbitrary structure.
- When minimizing χ^2 - like functions or functionals of more general case the linearization method can be used and in case of degeneration FUMIVI may automaticaly switch to accurate calculation of second derivatives matrix.

- Nonlinear Constraints of arbitrary structure can be used.
- Both analytical and numerical calculation of derivatives can be used.

The new code was extensively tested both on model and real data, the results of tests are given in section 3.

2. Algorithm

FUMIVI (as FUMILI) uses parameter restrictors, defining multidimensional parallelepiped ("box") around a point - current approximation. The minimized function on such a box is approximated by a quadratic function and at each iteration unlike FUMILI, its **approximate minimum** is searched for during one or a few substeps. An approximate quadratic function is built using the gradient and second derivatives either simplified as in FUMILI or full ones. If the quadratic function is positively defined, minimization does not put any serious problems. If, on the other hand, it is not positively defined, it can have many minima and there is no need to look for all of them during intermediate iterations. In principle, it is possible to calculate eigenvalues and eigenvectors of the matrix and then to learn the relief of the function and choose a reasonable direction.

Another method proved to be very effective. At each substep we modify a non-positivelydefined matrix so that it becomes weakly positively defined and in the valley direction a step becomes very big. The sides of the box prevent the movement outside of the box and the corresponding parameters are being temporarily fixed one by one. The rank of the matrix is reduced. The final matrix may become positively defined or minimum may occur in the corner of the box.

By the way, under consecutive fixing of parameters it is possible to decrease the order of inverse matrix by one without its full inversion. However it can be done only when the matrix became positively defined and wellconditionned because the lost accuracy cannot be returned. About the way of matrix regularization. During matrix inversion by the symmetric exclusion method we control the loss of accuracy and the sign of diagonal elements. The negative or equal to zero (at the limit of machine accuracy) diagonal elements are replaced by small positive ones, but not too small to avoid overflowing during final stage of matrix inversion. Another precaution is made while calculating the parameter step.

While searching for the minimum on the box we control the sum of squares of dimensionless constraint discrepancies (if the fit is done with constraints).

About movement along multidimensional vallies. For speedy movement along the crooked vallies we are using the following method. After we found the minimum of the approximate function on the box, we calculate the function value at the new point. If the function or the sum of dimensionless discrepancy squares decreased well enough then we move the box to a new point. Otherwise we are not hurrying to decrease the step - we recalculate a new approximate quadratic function at a new point and try to find minimum on the same, not moved box. As we mentioned before few parameters will be fixed because of box sides, conditionness of the problem may become better and we will have better chances to descend to the valley bottom, which is sometimes called firing a "crooked rifle".

When using the correct second derivatives in the case of narrow vallies we encountered an unpleasant case. It is easy to understand that while moving across a crooked valley the second derivatives matrix may change from negativelly defined on the inner side of the valley to a badly conditionned at the valley bottom and to a wellconditionned and positively defined on its outer side. In the movement along the crooked valley centrifugal force brings us to the outer side of the valley. In the outcome the matrix reciprocal to the second derivatives matrix becomes small and after multiplication by the gradient gives a small value of the parameter steps. The final result of this is slow movement along the outer side of the valley. A typical sign of this is the slow damping of the step though the step restrictors do not work. To cope with the situation it is possible to increase the step artificially but less risky to decrease the step restrictors in order to descend to the valley bottom after which the principle of the "crooked rifle" begins to work.

As we mentioned earlier one of the features of FUMIVI is automatic detection of situations when linearization method does not work. The long experience with FUMILI showed effectiveness of the linearization method in minimization of χ^2 - like functionals. So we saved this option for such types of functions. But when degeneration of the second type takes place the estimates of second derivatives in the step direction sharply change and their values are much less compared with accurate ones which can be estimated from function values. In these cases we switch off the linearization method and go to accurate calculation of second derivatives.

About constraints. As is well known popular package MINUIT[5] uses the variable metric method which does not require calculation of second derivatives. But if you want to take into account arbitrary constraints, it is not clear how to receive good convergence without calculation of second derivatives.

As is not well known, the necessary condition of the minimum of the regular function $F(X)$ under regular constraints $f_\lambda(X) \geq 0$ is the following : the gradient of the minimized function should expand into a linear combination of gradients of the active constraints (i.e. equal to zero) with nonnegative coefficients. Since for nonlinear constraints we cannot define the fact of correct equality to zero, we must introduce the conception of approximate equality to zero.

Remember that comparison of different values has sense when they have the same dimensionality. Used further down, the procedures of orthogonalization, sorting and selection of main element are not invariant to the change of scale. So we took decision to work in a dimensionless coordinate system and as the unity for each parameter we selected their error estimates if they exist. If not, we take the parameter restrictors as the unities. As the scale unity for discrepancy we take their formally calculated errors as functions of parameter errors(or parameter restrictors).

The logic of the work with the constraints is as follows. If inequality is satisfied and not equal to zero (in conception of approximate equality) we exclude such an inequality from consideration. Nonsatisfied inequalities are temporarily turned to equalities one by one in order to make move to the permitted region. The step restrictors, defining the box also take part together with inequalities but in case of noncompatibility have priority so that not to get out of box. In case of inequalities approximately equal to zero we do the following. First we select the number of such inequalities. Then, using the orthogonalization method we expand the gradient of minimized function $F(X)$ into a linear combination of their gradients and orthogonal addition. The constraints with negative expansion coefficients are excluded from consideration. Because in fact we minimize two functions we can investigate the components of the gradient of the sum of constraint squares. Then constraints are excluded only if the corresponding components of both gradients are negative.

So, before step calculation we may have three types of active constraints : first and of the highest priority - constraints, defining the box sides (approximately equal to zero). The next are crudely nonsatisfied inequalities and the last are those approximately equal to zero. The second group of equalities is sorted by constraint discrepancies. After forming such a system of linearized equalities we start the step calculation. In the very first equation we select the main element. Then we substract this equation from others so that to zero the coefficients, corresponding to the selected element. Then we select the main element in the second already transformed equation and so on. At the end we either find the formulae expressing the part of parameter steps over others or find the *linear dependence or incompatibility of equations. This fact is detected by controlling* the reduction of norms of transformed equation coefficients. If the reduction is more than some value for example 10^{10} (it depends on computer accuracy), we discard this equation. Because the equations are sorted by reduction of discrepancies, we discard the equation which is better satisfied. The typical cause of equation redundancy is activation of a big number of equations corresponding to box sides.

After we found substitution formulae we transform the quadratic form approximating the minimized function on the box. Then we analyze the second derivatives matrix, correct it if it is not positively defined and calculate the steps. However, when, as the result of such a step, the inequalities excluded earlier become nonsatisfied, we activate them (turning them into equalities) one by one, starting with the one with the biggest discrepancy. The activation of one inequality may lead to the situation when others become satisfied. Having at least one forcely activated inequality means that the minimum condition is not fulfilled. After getting final step we check if this step leeds to the decreasing of the sum of the squares of inequalities. If this sum increases because of too many equations discarded, we stop further movement in the box. If we had forcely activated inequalities we must repeat substep series from a new point with the same quadratic approximation of the function. But it is not practical to do them too many times. Generally speaking the most "honest" way is to express few parameters over others from the minimum condition for the sum of squares of all constraints. In particular it permits one to find some quasisolution when constraints are really incompatible. But for this we must also perform the substraction procedure described higher to do parameter subdivision. But by this moment we already have some solution and it is a pity to give it up.

Another important case is the case when in addition to the inequalities we have equalities. In this case it is necessary to mention that the equality $f_\lambda(X) = 0$ is equivalent to two inequalities $f_\lambda(X) \geq 0$ and $-f_\lambda((X) \geq 0$. In practice it means that while checking the minimum condition we must expand the function gradient in gradients of inequalities and gradients of equalities(both approximately equal to zero), but for equalities the analysis of signs of the corresponding components of the function gradient is not needed.

If minimization is going on succesfully the restrictors for which minimum is on the box side are increased four times. If on the contrary minimization process is rather difficult (as in FUMILI the precaution is made to avoid oscillations), all the restrictors are reduced (but not more than 8 times at a time). The iteration is considered to be successfull if we have good decrease of either the minimized function or the constraint discrepancies. The function may not decrease if there are nonsatisfied constraints. The requirement of necessary decrease of the discrepancies may lead to slow movement along curved bounds. Technically we introduce the sum of dimensionless constraint squares. Their scales are calculated at the beginning of iteration and comparison with the final value is made under the same scales.

The constraints may be linearized either simultaneously with the quadratization of minimized function or at each substep. The second variant is preferable if the function is undefined outside of the permitted region. In this case there is a possibility of satisfying constraints comparatively accurately at the end of every step. If the calculation of constraints is expensive, it is better to use the first variant.

In numerical calculation of derivatives as the natural differentiation step a small fraction (for example a hundredth) of restrictors may be used, at least before getting the final solution.

A few words about the control of the convergence. The end of the subiteration and iteration process may be controlled as in MINUIT by the value of expected function decrease. However under bad conditionnes even its sign can be wrong due to the accuracy loss. It is more reliable to compare steps with some fraction of error estimates as it is done in FUMILI. Here are also problems when one uses full second derivatives and has nonnegatively defined matrix during intermediate steps. Practice showed that good results take place when instead of parameter errors the small fraction of parameter restrictors is used. At the minimum the reasonable estimates of errors may be obtained.

One must remember that not only parameter steps must be small near the correct minimum, but crudely nonsatisfied or forcely activated constraints must be absent.

3. Tests

The new method, described here was developed as part of the software for an experiment on rare K^- decays [6]. It was extensively tested on model data for this experiment, in particular on $K_{\pi 2}$ - decays in topology, when momenta of the primary K-meson and both gamma rays are fully measured and only the direction vector of the secondary π -meson is measured. For such a case we have four constraints in the form of equalities

$$\begin{aligned} E_{\gamma 1} \cdot E_{\gamma 2} \cdot (1 - cos\theta_{\gamma\gamma}) - \frac{m_{\pi^0}^2}{2} &= 0 \\ E_{\gamma 1} \cdot [\mathbf{n}_\pi \times \mathbf{n}_{\gamma 1}]_x + E_{\gamma 2} \cdot [\mathbf{n}_\pi \times \mathbf{n}_{\gamma 2}]_x + p_K \cdot [\mathbf{n}_K \times \mathbf{n}_\pi]_x &= 0 \\ E_K - E_\pi - E_{\gamma 1} - E_{\gamma 2} &= 0 \\ E_{\gamma 1} \cdot [\mathbf{n}_\pi \times \mathbf{n}_{\gamma 1}]_y + E_{\gamma 2} \cdot [\mathbf{n}_\pi \times \mathbf{n}_{\gamma 2}]_y + P_K \cdot [\mathbf{n}_K \times \mathbf{n}_\pi]_y &= 0 \end{aligned} \tag{6}$$

one nonlinear inequaiity

$$P_\pi - (P_\pi)_{min} \geq 0 \tag{7}$$

and 8 linear constraints in the form $a \leq X \leq b$, corresponding to the necessary linear limitations on physical dimensions of the detectors. The conventions : $E_K, E_\pi, E_{\gamma 1}, E_{\gamma 2}$ - the energy of K-meson, π -meson and two gammas
$\mathbf{n}_K, \mathbf{n}_\pi, \mathbf{n}_{\gamma 1}, \mathbf{n}_{\gamma 2}$ - their directional vectors, normalized to unity.
P_K, P_π - momenta of K and π mesons.
We generated 100 events with the kinematics of this decay. Before the entry to the fit we "smeared" true values of parameters by the Gauss distribution with the errors ten times more than anticipated under real accuracies of our detectors. The measured coordinates themselves were not smeared. The idea was to get during the fit to the true values of the parameters, the total number of which was 14. We did not have a single failure of fit in

the sense that we always had convergence to some value. In 98 events of 100 we converged to the true values of the parameters, only in 2 we got wrong decision, but it is of common knowledge, because a solution depends on how you select the initial approximation for fitted parameters.

Another type of tests was done on real data with the same event topology. We analyzed about 150 events, taken practically without any preliminary selection. All of them converged to some solution and there was not a single nonconverged. For 3 events the number of iterations was more than 15.

4. Acknowledgements

We would like to thank Gurevitch M.I. who drew our attention to the correct form of the necessary condition of minimum.

References

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

[2] CERN Program Library, D510, FUMILI.

[3] A.J. Ketikian, E.V. Komissarov, V.S. Kurbatov, I.N. Silin, Nucl. Instr. and Meth. **A314**(1992)578.

[4] A.J. Ketikian, V.S. Kurbatov, I.N. Silin, Proceedings of CHEP-92, CERN 92-07, 1992,p.833

[5] F. James, M. Roos, CERN Program Library, MINUIT, D506

[6] V.N. Bolotov et al., Sov. J. Nucl. Phys. **45**,1023, translated from Yad. Fiz. **45**(1987),1652. from Yad. Fiz. **45**(1987),1652.

Received by Publishing Department
on July 2, 1993.