Fermi National Accelerator Laboratory

FERMILAB-Conf-89/172

Weighted Fit of Parametric Functions to Distributions - The New Interface of HBOOK with MINUIT *

Eliane S. Lessner
Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

August 1989

* Submitted to the 1989 Conference on Computing in High Energy Physics, Oxford, England, April 10-14, 1989.

Operated by Universities Research Association, Inc., under contract with the United States Department of Energy

Weighted Fit of Parametric Functions to Distributions–
The new interface of HBOOK with MINUIT.

Eliane S. Lessner
Fermi National Accelerator Laboratory[1]
Batavia, Illinois 60510.

Abstract

The fitting routines of the HBOOK package allow weighted fit of parametric functions to the contents of a one, two or N-dimensional distribution, and analysis of the function in the neighbourhood of its minimum, through an interface with the MINUIT package. These routines have been rewritten so as to interface the new version of MINUIT and to allow for smooth transitions to future versions of both packages. We discuss the interface and its capabilities: it is more stable than the previous version and presents a more accurate error analysis. The fitting algorithm is based on the Fletcher method, known for its reliability. Exponential, Gaussian and polynomial fitting are provided, as well as arbitrary user-defined fitting, to one, two and N-dimensional distributions. For the latter, the user is required to provide a smooth parametric function and is given the ability to guide the algorithm in finding the desired minimum. Examples are given.

1 Introduction

The interface of HBOOK [1] with MINUIT [2] has been rewritten to account for modifications introduced to the latter and to allow for smooth transitions to future versions of both packages and the interface of MINUIT with the PHYSICS ANALYSIS WORKSTATION system (PAW) [3].

HBOOK is a widely used subroutine package that handles statistical distributions (histograms and N-tuples) in a Fortran scientific environment. The MINUIT package, also widely used, performs minimization and analysis of the shape of a multi-parametric function in the neighbourhood of its minimum, where the function to be minimized is a chisquare or a log-likelihood function. The HBOOK fitting routines use the Fletcher minimization method, chosen for its reliability and stability. In particular, the MINUIT subroutine MIGRAD, which implements the method, has been rewritten to include a line search which improves the stability and accuracy of the covariance matrix. As a consequence, the analysis of errors is more accurate than that of previous versions [4].

The interface between the fitting routines and MINUIT takes full advantage of the new Fortran-callable mode of the latter and is contained in a single routine, which prepares the initial conditions (initial parameter values; step sizes; parameters' lower and upper bounds; etc.) in the format expected by MINUIT. This routine acts, therefore, as a "shell" between the two packages and can be easily modified to account for changes and/or new implementations.

In the following, we review briefly the Fletcher method; discuss the capabilities of the fitting routines and the chisquare calculation algorithm; and present some examples.

2 The Fletcher method

Variable metric methods, such as Fletcher's [5], are based on the differential properties of the function to be minimized over the parameter space whose metric is given by the matrix (G) of the

[1]Fermilab is operated by Universities Research Association, Inc., under contract with the U.S. Department of Energy

second derivatives of the function with respect to the parameters, the space coordinates. This formulation enables the construction of an invariant scalar: $EDM = g^T V g$, where: g is the gradient of the function with respect to the parameters and V is the covariance matrix. EDM represents twice the distance between the function value at the point where g and V are calculated and the minimum of a quadratic function with Hessian matrix G. EDM is, thus, an estimated distance to the minimum and provides the convergence criterion on which the method is based.
Minimization proceeds by taking quasi-Newton steps, where the step length is determined by quadratic interpolation and the search direction computed using a positive-definite approximation of the covariance matrix. After each iteration , the covariance matrix is updated according to a formula of the type: $V_1 = V_0 + F(V_0, \alpha_0, \alpha_1, g_0, g_1)$, where: α are the space coordinates, F is an appropriately chosen function , and subscripts 0 and 1 indicate values before and after the iteration, respectively. Convergence is achieved when some predetermined criterion on EDM is satisfied.

3 The fitting routines

The HBOOK fitting routines perform a weighted fit of a parametric function to the contents of a one,two or N-dimensional distribution and estimate the parameters by the least-square fitting method. The final values of the parameters correspond to those values for which the function:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{C(i) - F(x, a_1, \ldots, a_k)}{E(i)}\right)^2 \tag{1}$$

attains a minimum. Here, N denotes the number of channels (cells) in the one or two-dimensional histogram or number of points in the distribution; C(i) denotes the contents of channel or cell i; E(i), the error on C(i); x(i) is the coordinate of the center of channel (cell) i, or a coordinate vector of the point i, in a distribution; $a_1, \ldots, a_k$ are the parameters; and F is the theoretical parametric function.

Exponential (entry HFITEX); Gaussian (HFITGA); and polynomial (HFITPO) fitting are provided, as well as fitting to an arbitrary user-given function (entries HFITL and HFITS, for one or two dimensional histograms; and HFIT1 and HFITN, for non-equidistant points in a one-dimensional or N-dimensional space, respectively). For the last four entries, the user is required to provide a smooth parametric function. Notably, HFITL and HFITN offer the ability to guide the algorithm in finding the desired minimum, by allowing the choice of step sizes and parameter bounds deemed appropriate to the problem.

By option, the estimated distance to the minimum; number of calls; values of the parameters; standard deviations; and correlation coefficients are available after a predetermined number of iterations. Convergence to a minimum is achieved whenever EDM is less than 10E-4*the tolerance on the function value at the minimum (FEPS). By default, FEPS is set to 1.0, but the user has the option of choosing an appropriate value for his problem.
Abnormal termination occurs when either the number of iterations has been exceeded (default = 200), or convergence occurs at a point of non-zero gradient. In the first instance, the user has the option of increasing the number of iterations; in the second one, of either changing FEPS, or, probably better, changing the problem formulation, such as choosing different initial conditions (parameter values, step sizes or limits on parameters).
For the exponential; Gaussian and polynomial entries, the initial conditions are automatically derived from the histogram's statistical information and bin width. The remaining entries require the user to provide his own set of initial parameter values, given as an array. In particular, HFITL

and HFITN allow the user to select initial step sizes, lower and upper parameter bounds and are recommended for the analysis of more difficult problems, since they allow better control over the minimization algorithm.

A ***control word*** is provided in all entries, which allows selection of:

- superimposition the fitted function to the one-dimensional histogram.
- output printing level, from no output to output at every each selected number of iterations.
- weights, which can be either set to 1.0, or calculated as the square root of the contents, unless the one-dimensional histogram is weighted with errors computed at filling time.
- numerical calculation of derivatives or analytical (user-given) calculation of derivatives (requires user routine).This latter choice speeds up the computation time, when the user chooses to force the algorithm to accept his calculations. Further, a new option exists of letting the algorithm accept these values only when their comparison to the program's finite-difference calculated values is acceptable according to some criterion. We find this option particularly useful for difficult problems, since the Fletcher algorithm depends crucially on the accuracy of the first derivatives, and "fails miserably" if they are not accurate. The cost involved is, of course, an increase in computing time (in some cases, by a factor of four).

Precision: meaningful results can only be achieved if the essential computations are performed with a sufficient wordlength. On VAX, IBM and APOLLO mainframes, the standard fitting routines require DOUBLE PRECISION and the parametric function, as well as input parameters must be declared DOUBLE PRECISION; idem, for the user-supplied derivatives, when applicable.

4 Computation of chisquare

The algorithm that calculates the distance between the theoretical function and the distribution has been rewritten. Previously, the algorithm ignored channels or cells with no entries, defined as those channels (cells) for which the content, divided by the distribution's maximum weight, is smaller than 1.E-10. However, the known problem of fitting with a small number of entries per bin [2] led us to include all bins whose errors are non-negative. Unless errors have been assigned at filling time, the contents and errors are summed partially, until a nominal test value (square sum of contents divided by the square sum of errors) is greater than 1.5. The summed bins are counted as one point used in the fit (one degree of freedom). Tests comparing the fit obtained by not summing and by summing small entry bins show that a better fit is achieved when the latter procedure is effected.

5 Examples

Figure 1. displays the fitting of a two-superimposed Gaussian to a histogram that has been filled with random numbers distributed according to the contents of a histogram generated by the values of the same two-Gaussian function. The interesting feature of this example is the existence of, at

[2]The Gaussian error of a low fluctuation is too small, so that the chisquare fit may give a total area under the function smaller than the area of the histogram [6].

least, two minima, the "physical" one, which we seek, and another, for which the parameters take unphysical values, with much higher chisquare.

For this figure, the fit is done *not* summing bins with small number of entries.

The function to be superimposed has the form:

$$F = C1e^{-\frac{1}{2}(\frac{X-XM1}{XS1})^2} + C2e^{-\frac{1}{2}(\frac{X-XM1}{XM2})^2} \tag{2}$$

The histogram to be fitted is booked with 100 channels, ranging from 0.0 to 1.0, with a maximum population of 1000.0 entries per channel. Based on the random-generated histogram, we assign initial values C1=200.0, C2=100.0 (corresponding roughly to the histogram peaks), XM1=0.4, XM2=0.6 (positions of the peaks) and XS1=0.1, XS2=0.1 (deviations), with all parameters unbounded, and let the algorithm (HFITS) choose initial step sizes. Weights are taken as the square root of the contents and the derivatives are calculated numerically.

The final reduced chisquare is 27.13. Note the final value for the parameter XM1=-0.24913, which is obviously unphysical. Note, also, the large errors on XM1 and XS1.

Figure 2. displays the fit achieved with the same initial values as given previously, assigning an error of 1.0 to empty bins and summing bins of small entries. The final reduced chisquare is 1.25. All the final values of parameters are positive and close to the expected, with small errors.

6 Conclusion

The new HBOOK fitting routines provide flexible and reliable algorithms for fitting of distributions, made more accurate and stable by its interface to the new version of MINUIT and by modification of the chisquare computation. Forseen implementations are algorithms for N-tuple fitting and the integration of the full capabilities of MINUIT to the PAW system.

Acknowledgements

The author is indebted to Dr. Frederick James and Dr. Rene Brun for many useful discussions.

References

[1] R. Brun, D. Lienart, HBOOK User Guide, *CERN Computer Center Program Library Y250*, October, 1987.

[2] F. James, M. Roos, MINUIT Function Minimization and Error Analysis, *CERN Program Library D506*, April 1989.

[3] R. Brun, O. Couet, N. Cremel, C. Vandoni, P. Zanarini, PAW – Physics Analysis Workstation, *CERN Program Library Q121*, September, 1988.

[4] E. Lessner, Fermilab. Case study, unpublished.

[5] R. Fletcher, A New Aproach to Variable Metric Methods, *Comput. J. 13, 317*, 1967.

[6] I. C. Brock, MN–FIT. A Fitting and Plotting Package using MINUIT, *CERN*, August, 1988.

```
HBOOK        ID =        3                                            PARAMETER     VALUE         ERROR
                                                                         C1         63.901        2.4591
     180                                -                                C2         40.095        1.0138
     170                                I  -                             XM1       -0.24913       0.21580
     160                               -I  I--                           XM2        0.54819       0.49256E-02
     150                             - II -I I                           XS1        0.58058E-01   0.13553
     140                             I II-I  I                           XS2        0.23260       0.29780E-02
     130                             I-I     I-
     120                             I        I                       REDUCED CHISQUARE  = 0.2713E+02
     110                           - I        I-
     100                           I-I         I---
      90                          -I              I                         --    -    -
      80                          I               I                         II -- I -- I-
      70                        --I               I                       - II-II-I-II II   -
      60                        I                 I--                 --- -I-I        I-II---I
      50                       -I                   I****************************           I --
      40                     - I            *********-I-          --I            ********   I-II- -
      30                    -I-I    ********           I      ----I                      *********I----
      20                - **********                   I------I                                   **********
      10     *************I                                                                             I--I****

CHANNELS 100 0                                                                                                  1
          10 0        1         2         3         4         5         6         7         8         9         0
           1 1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890

CONTENTS 100                        1 11111111111
          10             1 112334668094257346552199955343211111222233455545758877768777587565634432323222111 1
           1.      21 252163123026476638679147750416771330875895569589315930913043440397234034641983204281073231

LOW-EDGE  1.           111111111122222222223333333333444444444455555555556666666666777777777788888888889999999999
*10** 1  0   0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789

* ENTRIES =        200      * ALL CHANNELS = 0.5000E+04      * UNDERFLOW = 0.0000E+00      * OVERFLOW = 0.0000E+00
* BIN WID = 0.1000E-01      * MEAN VALUE   = 0.4821E+00      * R . M . S = 0.2180E+00
* CHISQUAR  =  0.2333E+04
```

FIGURE 1. Fitting of two superimposed Gaussians, by subroutine HFITS, not summing bins with small number of events.

```
     HBOOK     ID =         3                                               PARAMETER       VALUE        ERROR
                                                                               C1           159.01       4.0020
                                                                               C2           78.160       2.0267
     180                                   -                                   XM1          0.30129      0.12663E-02
     170                                   I  -                                XM2          0.69567      0.25739E-02
     160                                  -****--                              XS1          0.67040E-01  0.10827E-02
     150                                - *I -I**                              XS2          0.11393      0.20987E-02
     140                                I*II-I  I
     130                                *-I     I*                          REDUCED CHISQUARE = 0.1258+01
     120                               *I        I*
     110                              - I        I-*
     100                              *-I         I---
      90                             *I             *I                         --    -    -
      80                             I               *                         ***********I-
      70                           -*I               I*                     ***II-II-I-II ***  -
      60                           *                 I-*                --**-I-I        I-II-**I
      50                          *I                  I*-            ***I-I                  ***-
      40                        -*I                   I-*-        **I                       I-I** -
      30                       **-I                      **     ****I                           I***---
      20                   - **I                         I-*****I                                   *****  -
      10       **************I                                                                          I*******

CHANNELS 100   0                                                                                                  1
          10   0        1         2         3         4         5         6         7         8         9         0
           1   1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890

CONTENTS 100                          1 11111111111
          10               1 112334668094257346552199955343211111222233455545758877768777587565634432323222111 1
           1.        21 252163123026476638679147750416771330875895569589315930913043440397234034641983204281073231

LOW-EDGE   1.            111111111122222222223333333333444444444455555555556666666666777777777788888888889999999999
*10**  1   0   0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789

* ENTRIES =       200     * ALL CHANNELS = 0.5000E+04     * UNDERFLOW = 0.0000E+00     * OVERFLOW = 0.0000E+00
* BIN WID = 0.1000E-01    * MEAN VALUE   = 0.4821E+00     * R . M . S = 0.2180E+00
* CHISQUAR  =  0.1013E+03
```

FIGURE 2. Fitting of two superimposed Gaussians, by subroutine HFITS, summing bins with small number of events.