Auto-optimizing Gaussian Fitter

Intel Science Talent Search
(Physics)
November 2012

William Se Hwan Kim 김세환

Phillips Academy Andover
University of Texas at Arlington
High Energy Physics Group

Abstract:

This paper explains the algorithm for the Auto-optimizing Gaussian Fitter and also serves as a user guide for the program. This algorithm is used for fitting data for Gas Electron Multiplier (GEM), a detector candidate for the next generation International Linear Collider. The GEM chambers use Gaussian Fitting method to statistically analyze the detected particle data, and this Auto-Fit code automatically seeks the fit range while preserving its accuracy. This program not only reduces each data set's processing time, but also offers a batch analysis function, which processes data in large quantities.

Contents

1. Introduction

Scientists and engineers use experimental data to create new theorems and conjectures. Finding a model for the raw data is very important. The process of finding the appropriate mathematical model proves to be a complex task, also requires a strong calculational firepower. The advent of computers has allowed scientists and engineers to rely on computer to fit data, and many statistical analysis softwares have been invented to simply the process.

Over the summer, I worked for the High Energy physics group of University of Texas at Arlington. This group conducts research on Gas Electron Multiplier (GEM), which is a strong detector candidate for the upcoming International Linear Collider [MY]. The University of Texas at Arlington's High Energy Physics group has developed prototype GEM chambers, and is currently conducting experiments to understand these chamber's characteristics. The GEM chambers collect data and recorded as a digital signal. The raw data requires fitting in order to yield the measured signal. It is important to fit the data in a correct range while keeping the accuracy to a certain level. Because the chamber measures charges in a scale of femto Coulombs (fC), the fitting must also keep its precision and accuracy. However, when dealing with raw data like that of a GEM chamber, finding the optimal fit is a tough task, as it has no specific method and varies data set by data set. The single detector in a collider creates an astronomical amount of data sets, as a result, the speed of processing this data becomes crucial.

I coded a program in Java language that auto-fits the raw data sets. The Auto-Fitting method reduces the processing time for this delicate data set while maintaining the precision. This Auto-Fitting technique uses Chisquare and the degree of freedom to navigate the optimal fit range. This paper will analyze the auto-fit, and the use of this program.

2. Fitting Method

Searching for the model that can best describe a data set can be a very arduous and complex task. Nonetheless, finding this correct model can be very helpful in estimating important characteristics of the given data set such as the rate of change on the curve, the minimum and maximum values of the function, and prediction for other values. The process of Fitting data plays an important role.

This process entails fitting the data to pre-defined functions, and its goal is to find the optimal parameter that closely models the data. The main functions involved in this process are linear regression and curve fitting. Here are the two main functions.

2.1. Linear Regression

Fitting consists of two models [1]. The first one models linear regression [1]. This model is used when the relationship between the two variables is linear [1]. Here is the equation:

$$y = a + bx$$

Here, the x is an independent variable, and y is the dependent variable [1]. The constant a is the y-intercept and b is the slope of the function [1]. Linear regression aims to find a and b that creates a function closest to the actual data [1].

2.2. Nonlinear Curve Fitting

In most cases, the data proves to be more complex than a linear regression [1]. Complex curves are better described with a nonlinear fitting model [1]. When using a nonlinear model, fitting requires a minimizing method [1]. The minimizing method that is mostly widely used is the Levenberg-Marquardt Algorithm [1]. This method is also used in the Auto-Fit Gaussian Program.

2.3. Levenberg-Marquardt Algorithm

The Levenberg-Marquardt Algorithm is an iterative algorithm used to minimize the following function [1]. This is the Levenberg-Marquardt algorithm aims to minimize this equation:

$$\chi^2(a,b) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i,a,b)}{\sigma_i}\right]^2$$

The algorithm requires initial parameters [1]. By giving variation to these parameters, the algorithm attempts to minimize χ^2 [1]. Here, function $f(x_i,a,b)$ is the fit function applied to the data to [1]. χ^2 and x_i represents the x and y coordinates of the data points [1].

2.4. Evaluating Fit

For a given fit curve, there is a way to assess whether the curve describe the raw data well. Chisquare and degree of freedom is used to measures how well the fit data matches the actual data [1]. These provide a quantitative value, and is also called the standard deviation of the residual:

$$\frac{\chi^2}{ndf}$$

Here, the closer this value approaches 1 the better fit is [1]. The AutoFit Gaussian program uses this chisquare divided by degree of freedom value to search the optimal fit range for the fit function.

3. Auto-Fit Code/Algorithm

The Auto-Fit Code utilizes the `if-else` and `while` loops to perform the algorithm. It receives the variable fileNum for each given data file.

```
if(isAutoFit){
	while(true){
		fittingList.get(fileNum).plusNumberOfIterations();
		double nowNdf=fitting(fileNum);
		if(Math.abs(nowNdf-1)>=Math.abs(beforeNdf-1)){
			break;
		}
		beforeNdf=nowNdf;
	}
	System.out.println("in AutoFit");
}
else	{
	System.out.println("in Not AutoFit");
	fittingList.get(fileNum).plusNumberOfIterations();
	fitting(fileNum);
	fittingList.get(fileNum).setFitRangeMin(minFit);
	fittingList.get(fileNum).setFitRangeMax(maxFit);
```

Figure 1. Auto-Fit Code.

The `isAutoFit` is a boolean operator, and once the Auto-Fit program initiates, the operator changes to true. The following `while loop` serves as the main auto-fit algorithm. The `fitting(fileNum)` returns the value of Chisquare/ndf. This value will be the key for navigating the optimal fit range.

The next if loop statement `(Math.abs(nowNdf-1)>=Math.abs(beforeNdf-1))` is the exit condition. If the previous Chisquare/ndf value is closer to 1 than the current Chisquare/ndf value, the loop breaks. This is when the Auto-Fit algorithm ends.

The code outputs the file in the same directory as the batch file.

4. Program

4.1. GUI

Figure 2 is the Graphic User Interface of this program. The first part for the Histogram has fields where one can input Histogram information. The first field “Data File Input” requires raw data in order to create histogram. The program can accept any text file (.txt). The next seven fields are for basic parameters for creating histograms. After filling in this basic information, one can create a histogram by pressing the Histogram button.

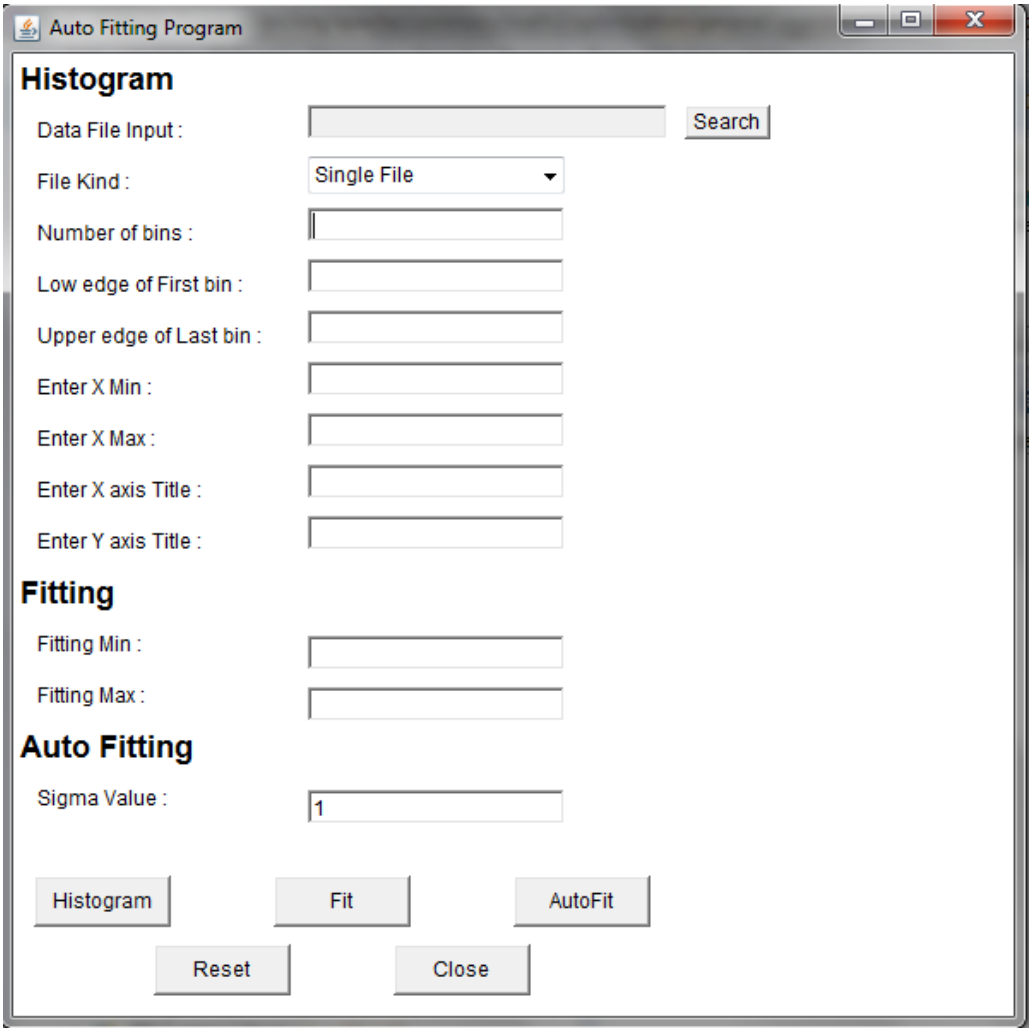

Figure 2. Graphic User Interface of the program.

Figure 3 shows a model histogram. The x and y labels can be entered through the previous form. The last three fields of the form set the parameters for the fitting method. Currently, the only fitting method supported is Gaussian Fitting; however, I plan to add more fitting methods. The remaining two fields set the minimum and maximum of the fitting range. The last field is for the auto-fit method. The user can choose the width of the fit data, by entering the sigma value.

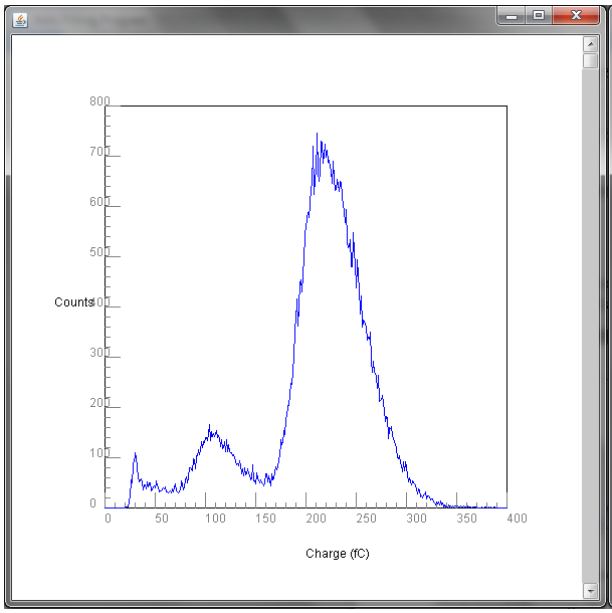

Figure 3. Example Histogram.

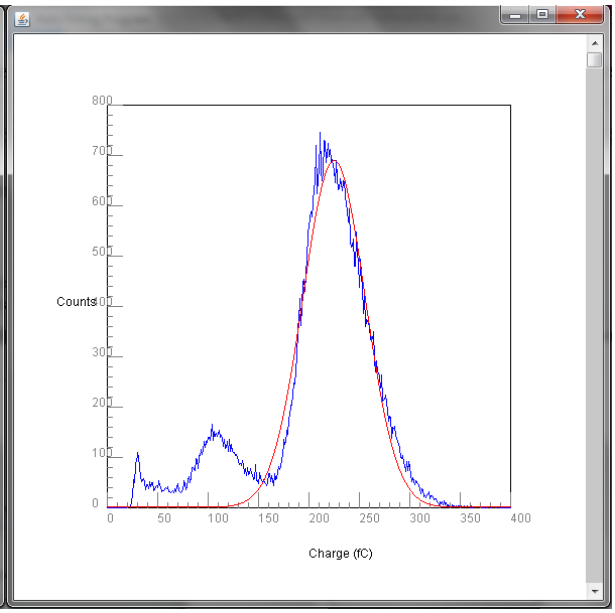

Figure 4. Fit range from 0 to 400.

Figure 4. is an example of data fit with ranging from 0 to 400. The red line shows the fitted function. The contrast between the red and the blue lines shows how well the fitted function fits the histogram.

The console panel lists the optimal parameters for the minimized function.

```
mConstant:688.6423481737617
mFitMean:224.7954768227473
mSigma:30.85795564465148
startX:193.0,endX:255.0
sum:242.68463224566239
ndf:61.0
ndfTest:3.978436594191187
numberOfIterations:1
data count:63161
```

Figure 5. Console Panel.

The Auto-Fit button then auto-fits the function for the given sigma value. The sigma value with a default of 1 can be changed by the user. By changing the sigma value, the user selects the width of the data for analysis. The example below displays a 1 sigma auto fit. Note that this curve provides a better fit than the curve from figure.

5. Data Sets

5.1. Controlling Data Width

By changing the last field labeled sigma, the user can control the width of the data . The sigma value is not limited to integer values only. User can enter any number that is under 6.

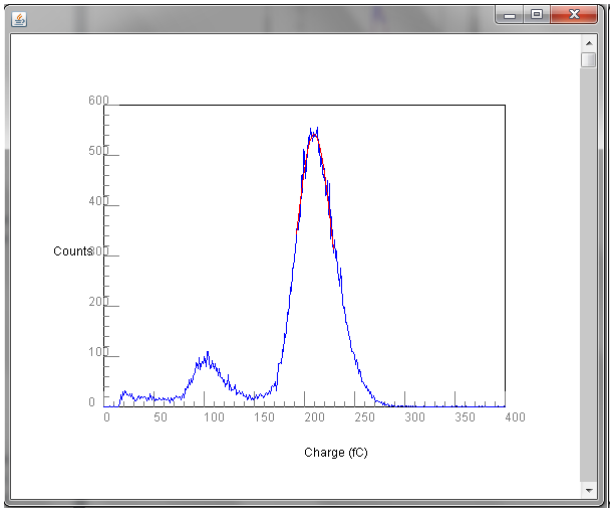

Figure 6. Fit data with 1 Sigma.

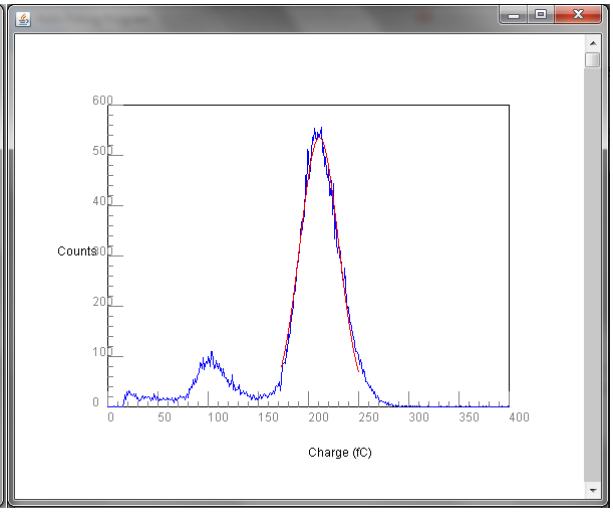

Figure 7. Fit data with 2 Sigma.

5.2. Single File

For single file, the user can input one .txt file that has the raw data. Here is the input data file. The data in this file is the charge measurement from a 3x3 GEM chamber. This data is a Fe_{55} radioactive source:

```
191.9347
246.1891
86.9411
281.5345
237.0982
136.0561
235.2558
246.3588
228.8558
260.7103
270.2375
248.1527
271.5709
297.7042
```

The data file must be listing the raw data. And once the user presses the Auto Fit button, the result will be :

5.3. Batch File

This program allows the user to process data in batch. The user has to create a batch file with the list of the data files. The data files and the batch list file should be in the same directory. The batch list file should look like this:

```
20120703-7.dat
20120703-8.dat
20120703-9.dat
```

Once the user create a batch list file, the user must change the drop box field as batch file from the initial setup, single file. After the user hits the Auto-Fit button, the program output file will be created in the same directory as the input file with the file name: `AutoFitting.txt`. The batch files will not display all the histograms.

```
--20120703-7.dat--
Entries:63352
default Mean:167.84292524308626
Sigma:18.709113631806837
Fit Mean:184.60996073270746
Fit Range Min:165.0
Fit Range Max:203.0
chisquare/ndf:37.30280746274021/37.0
Constant:992.7918595767926
number of iterations:5


--20120703-8.dat--
Entries:51062
default Mean:171.63849829618894
Sigma:18.991080225023786
Fit Mean:185.57930693359242
Fit Range Min:166.0
Fit Range Max:204.0
chisquare/ndf:40.218462914234465/37.0
Constant:806.3702055959177
number of iterations:5


--20120703-9.dat--
Entries:58266
default Mean:169.20138674355542
Sigma:19.343084968381632
Fit Mean:186.15678146289753
Fit Range Min:166.0
Fit Range Max:205.0
chisquare/ndf:41.9331239012783/38.0
Constant:891.8853688815574
number of iterations:5
```

Figure 8. Output File for batch file.

Figure 8 shows an example output file. The result comes out in a format of entries, mean, sigma value, fit mean, fit minimum, fit maximum, chisquare/ndf value, and the number of iterations.

6. Conclusion

Modern experimental science strongly relies on computing. Stronger computational power and code efficiency proves to be key as it can reduce time and computational load when analyzing massive amounts of data sets. The Auto-Fit program provides an efficient method before choosing the optimal fit range, always a tedious and time-consuming step in analyzing data. Also, the function that allows batch data analysis is much convenient for particle physicists who must analyze data on a large big scale. Instead of applying the fitting method one-by-one, this program can process all the data at once.

Most statistical analysis software provides functions for creating histograms and various fit methods, but not the useful functions like an Auto-Fit method. The Root program, Origin from Oracle, and the Experimental Data Analyst (EDA) are the three most popular tools for experimental scientists. Though they have strong graphic user interfaces, they still lack efficient tools for statistical analysis. However, the Auto-Fit program only processes data that specifically data that can be analyzed Gaussian fit method. For future work, I plan to expand and add to this code, adding several other fitting methods.

This program will be used by the high energy physics group of the University of Texas at Arlington. If the GEM chamber is chosen as the preferred detector candidate for the upcoming International Linear Collider, this program will be a crucial to the functioning of these chambers.

7. Bibliography

[1] Ledvij, Marko. http://www.originlab.com/index.aspx?go=Products/Origin/DataAnalysis/CurveFitting&pid=967

[2] Park, S., Balelomar, E., Measurement of Gas Electron Multiplier (GEM) Detector