Statistical Data Analysis, Assignment 5

Roel Deckers
Utrecht University

October 9, 2018

1 Objective

In this report we try to fit a model of the form

$$M_{a,b}(m) = \frac{\mathrm{d}N(m)}{\mathrm{d}m} = a \times m + b, \tag{1}$$

with model parameters a and b to a histogram of a mass distribution of events detected in a collider experiment using the maximum likelihood method.

2 Theory

Given histogram with bin centers $x_0, \ldots, x_n$ and corresponding counts $C_1, \ldots, C_n$, the likelihood $L_{M_{a,b}}$ of model $M_{a,b}(m)$ is given by

$$L_{M_{a,b}} = \prod_{i=0}^{n} \mathrm{Poisson}(C_i; \mu = M_{a,b}(x_i)). \tag{2}$$

That is, the product over all bins of the odds of seeing a measured bin count C_i, assuming measured bin counts are distributed according to a Poisson distribution with a mean given by the model prediction at the bin center.

Taking into account the definition of the Poisson distribution

$$\mathrm{Poisson}(k; \mu) = \frac{\mu^k e^{-\mu}}{k!}, \tag{3}$$

and writing the factorial in terms of the Γ-function, the log-likelihood can be written as:

$$\log L_{M_{a,b}} = \sum_{i=0}^{n} \left(C_i \log(M_{a,b}(x_i)) - M_{a,b}(x_i) - \log\Gamma(C_i + 1) \right). \tag{4}$$

Which will be the quantity we try to maximize in order to determine a maximum likelihood fit. We use $\log\Gamma(k+1)$ because direct computation of $k!$ is computationaly infeasible for larger k and there exist[1] good premade approximations to $\log\Gamma$.

[1] `https://root.cern.ch/root/html524/TMath.html#TMath:LnGamma`

3 Data

Our to-be-fitted histogram is shown in figure 1. We note that the data is restricted to the exclusive range $(1, 3)$, therefore we posit that as an additional restriction on our model we have $M_{a,b}(m) > 0 \, \forall \, m \in [1, 3]$ such that there is a non-zero change of seeing any value in $(1, 3)$ in the dataset.

In order to restrict ourselves to this domain we reformulate our model parameters from a slope a and bias b to y_1 and y_3: the values of M at $m = 1$ and $m = 3$ respectively. In these coordinates the non-zero restriction simplifies to $y_1 > 0$, $y_3 > 0$. In order to recover a, b from y_1, y_3 one can use the formulae

$$a = (y_3 - y_1)/2, \tag{5}$$

$$b = (3y_1 - y_3)/2. \tag{6}$$

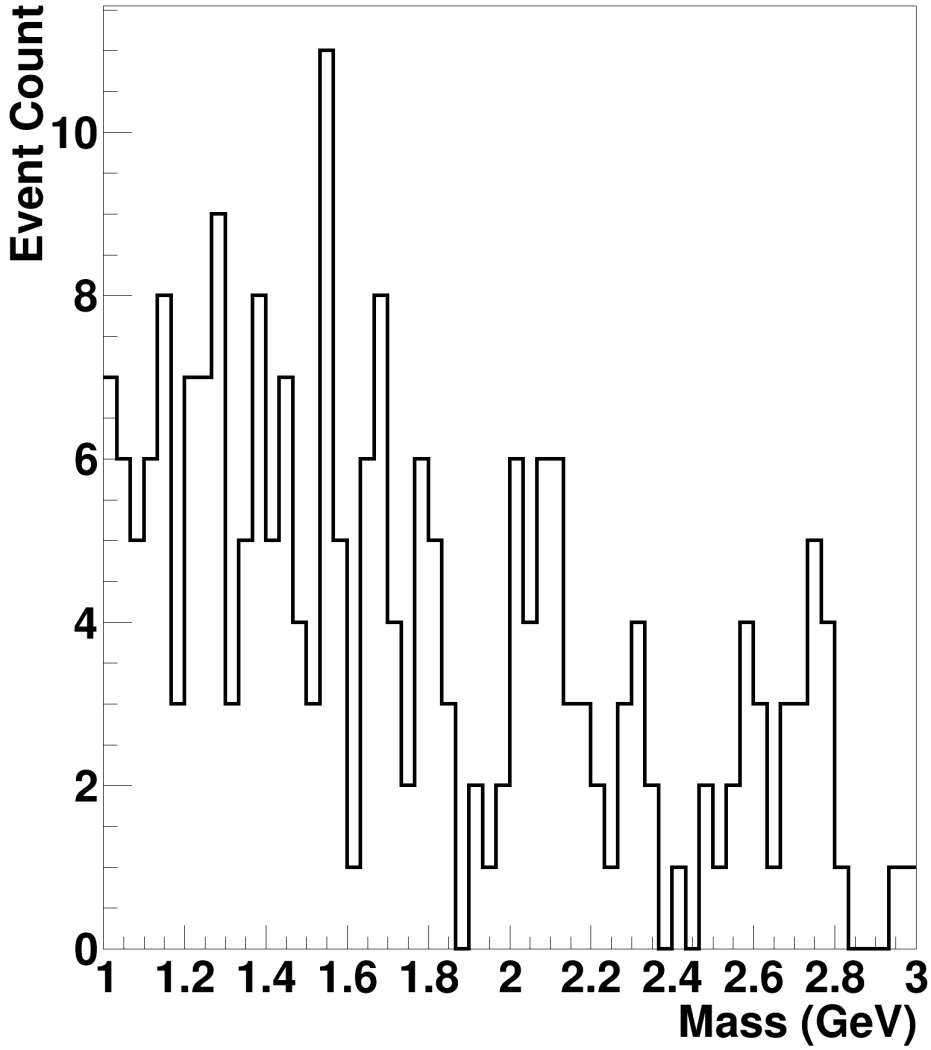

Figure 1: The to-be-fitted original data.

4 Constant Model

We start our analysis by fitting the model Ma=0, $b = M_{y1=y2=b}$ to our data, that is: we fit a horizontal line. For this we apply an iterative Newton solver to optimize the function $-\log L_{M_{y1=y2=b}}$. We find that the optimal value $\hat{b}$ is given by:

$$\hat{b} = 111 \pm 7. \tag{7}$$

The result is plotted in figure 2.

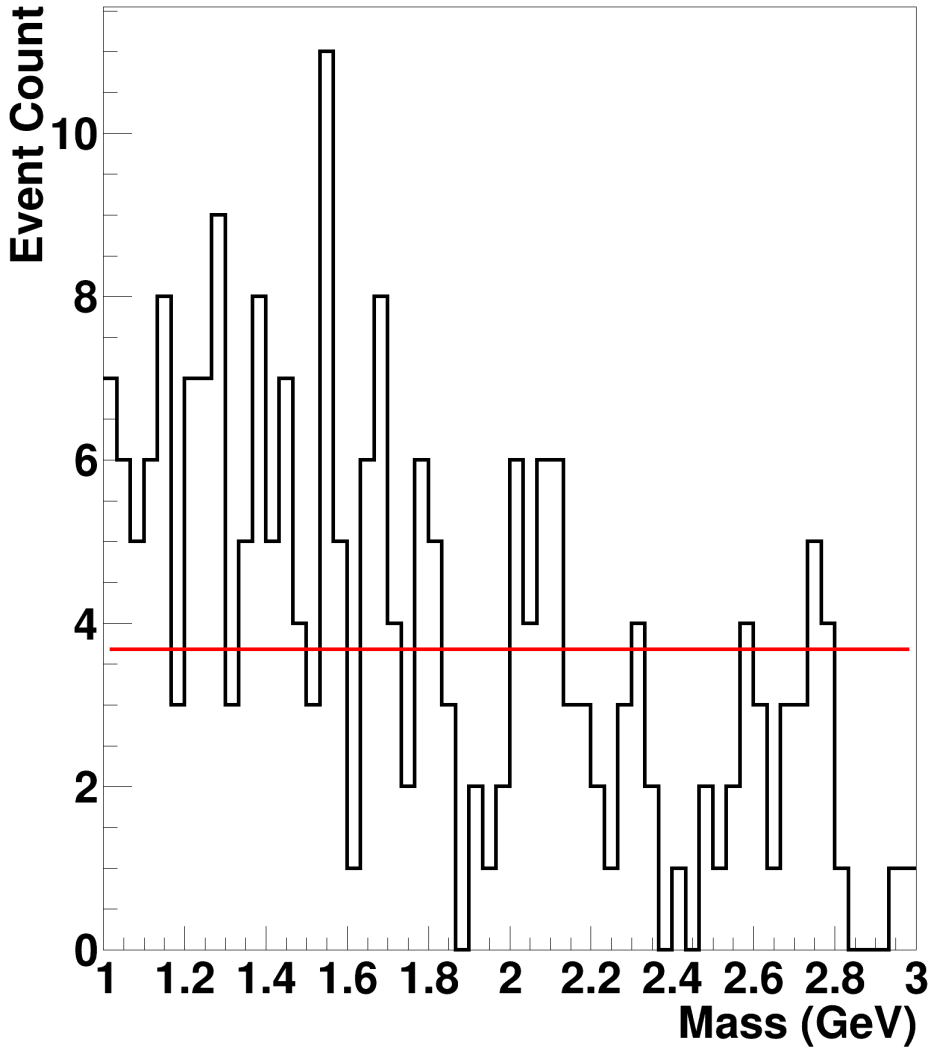

Figure 2: Fit of the constant model $M_{a=0,b} = M_{y1=y2=b}$ to the original data. The red line is the optimal model given by $\hat{b} = 111 \pm 7$.

5 Linear Model

We now refine our analysis by fitting the complete model $M_{a,b}$ to our data, that is: we fit a linear model. As noted in the Theory section, the model $M_{a,b}$ has a less desirable domain

compared to M_{y_1,y_3}, however we know that the model $M_{a=0,b=111}$ is a valid model and we posit that starting from this point, our optimizer will converge to the global minima without trouble.[2] Indeed it does, it quickly converges to

$$\hat{a} = -85, \tag{8}$$

$$\hat{b} = 280, \tag{9}$$

with covariances (as determined by the inverse of the numerical Hessian)

$$\sigma_{a,a} = 134, \tag{10}$$

$$\sigma_{b,b} = 761, \tag{11}$$

$$\sigma_{a,b} = -311. \tag{12}$$

These results are shown graphically in figure 3 and figure 4. In these figures one can see the objective function versus a an b alongside the path taken by the Newton optimizer (black & red) as well as the 1-σ contour of the final result (green). The final fitted model is shown in figure 5.

6 3rd-degree Polynomial model

We now introduce another model, that of a third degree Polynomial:

$$M_{a,b,c,d}(m) = \frac{\mathrm{d}N(m)}{\mathrm{d}m} = a + m\left(b + m\left(c + md\right)\right). \tag{13}$$

When we fit this model we find the optimal values

$$\hat{a} = 439, \tag{14}$$

$$\hat{b} = -312, \tag{15}$$

$$\hat{c} = 101, \tag{16}$$

$$\hat{d} = -14. \tag{17}$$

These results are presented graphically in figure 6. While this model has twice the degrees of freedom as our old model has. A close inspection shows that it hardly deviates from the from the previous model. A proper ^{2}-analysis would almost certainly show that this model is less likely than the linear model.

7 conclusion

We speculate that the $M_{a,b}$ model with parameters $\hat{a} = -85, \hat{b} = 280$ is the best model shown. A lower degree model did not adequately model the distribution while a higher degree model did little to improve the fit. A more rigorous analysis remains to truly confirm these speculations.

[2] In general a Newton Optimizer could end up outside of the domain even when starting near the minimum due to the parameter γ, however due to a quirk in our implementation this is not an issue: results outside of the domain of the model return 'NaN' which will is caught by the optimizer as a sign that γ is too big anyways. By working in a, b space we can report the covariances directly without having to convert between coordinates :)

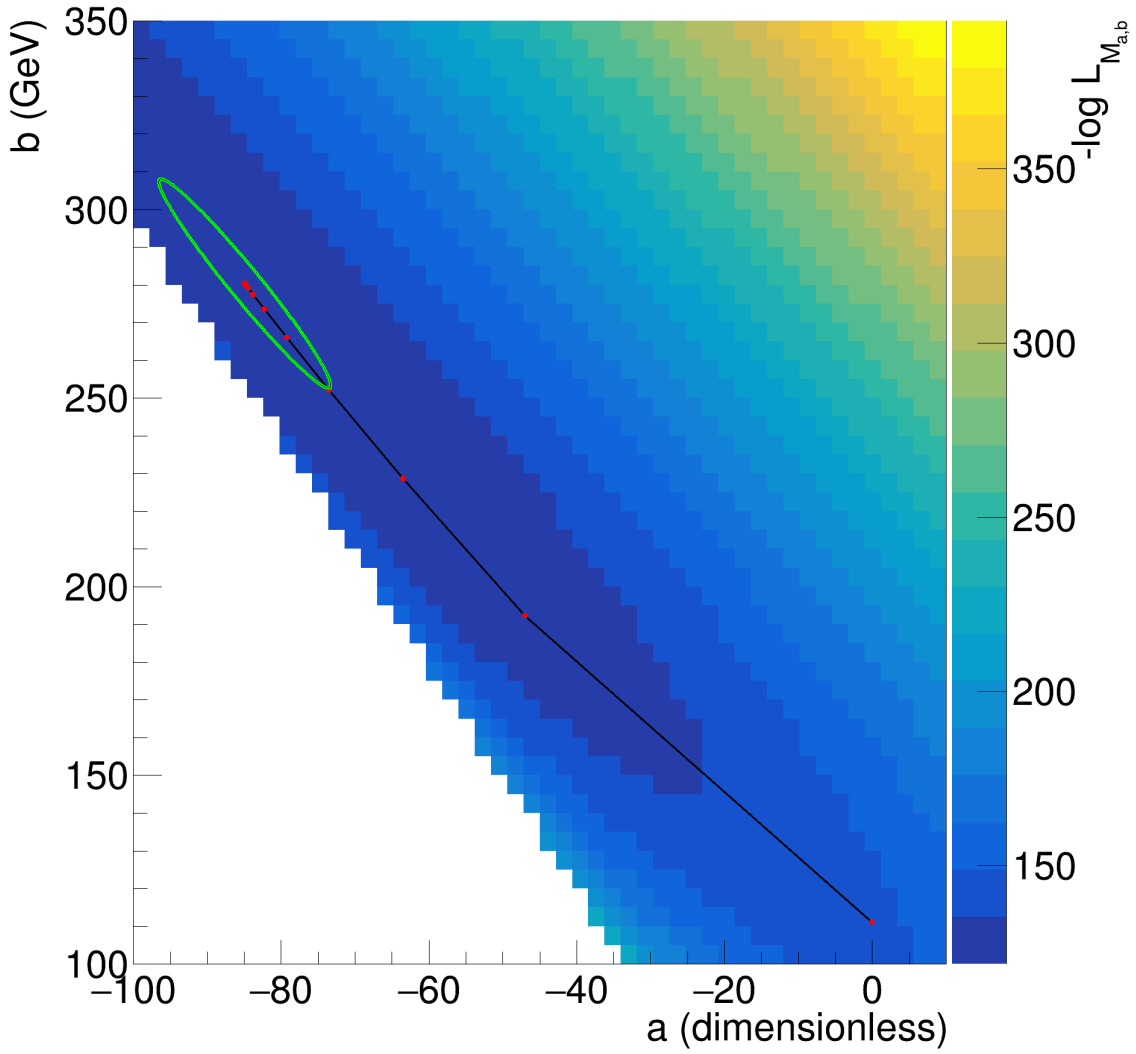

Figure 3: Fit of the linear model $M_{a,b}$ to the original data. The red and black line is the path followed by Newton optimizer towards the optimal values of $\hat{a} = -85, \hat{b} = 280$. The green curve is the 1-σ contour line of the optimal point. Note that values outside of parameter space of $M_{a,b}$ are shown in white.

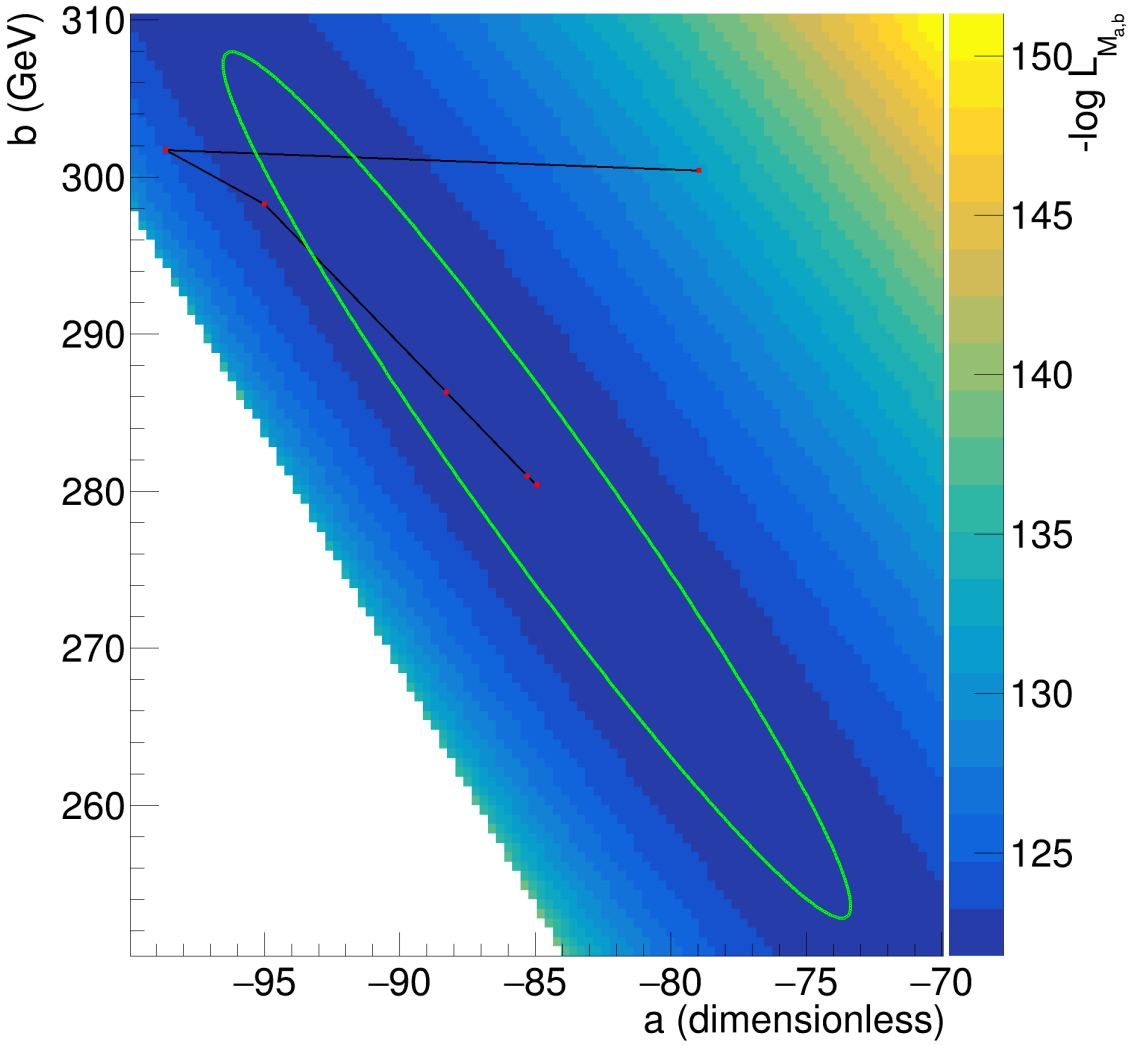

Figure 4: Fit of the linear model $M_{a,b}$ to the original data. As in figure 3 only zoomed in closer to the optimal values and with the optimizer started from a different initial value.

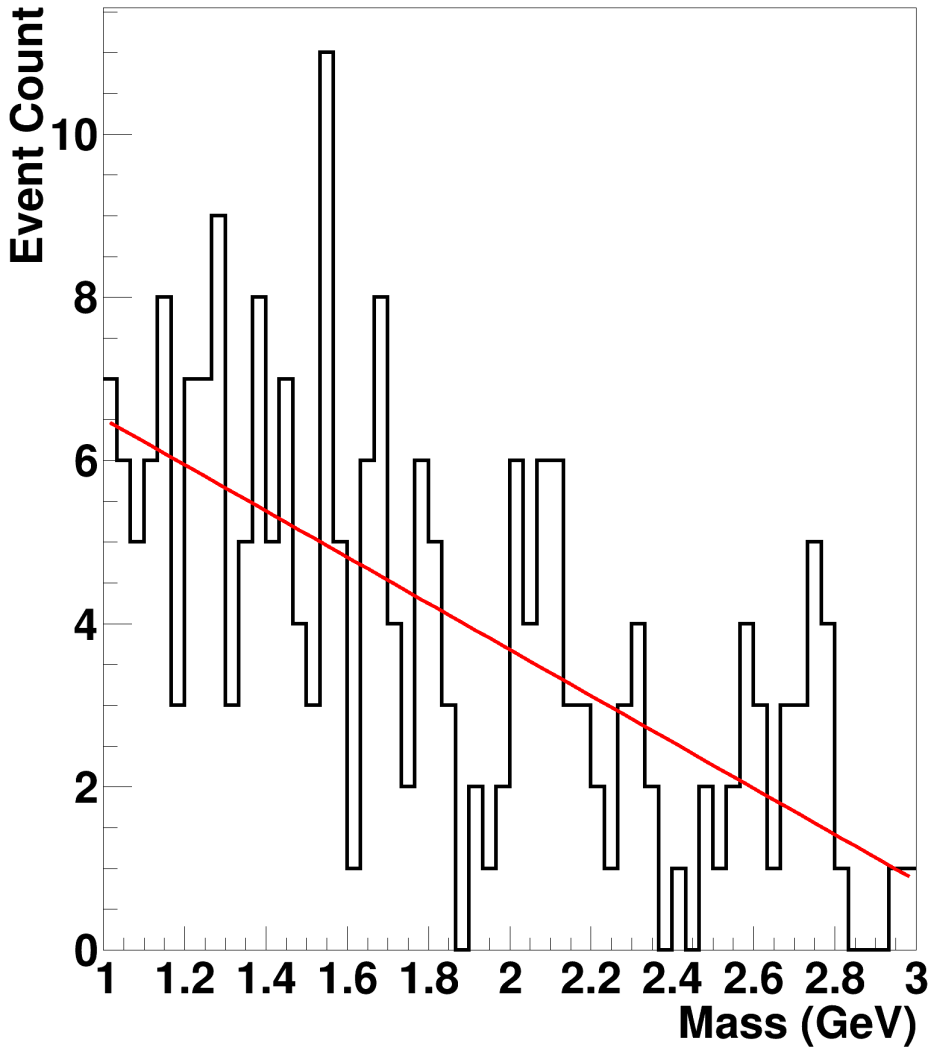

Figure 5: Fit of the linear model $M_{a,b}$ to the original data for optimal values $\hat{a} = -85, \hat{b} = 280$.

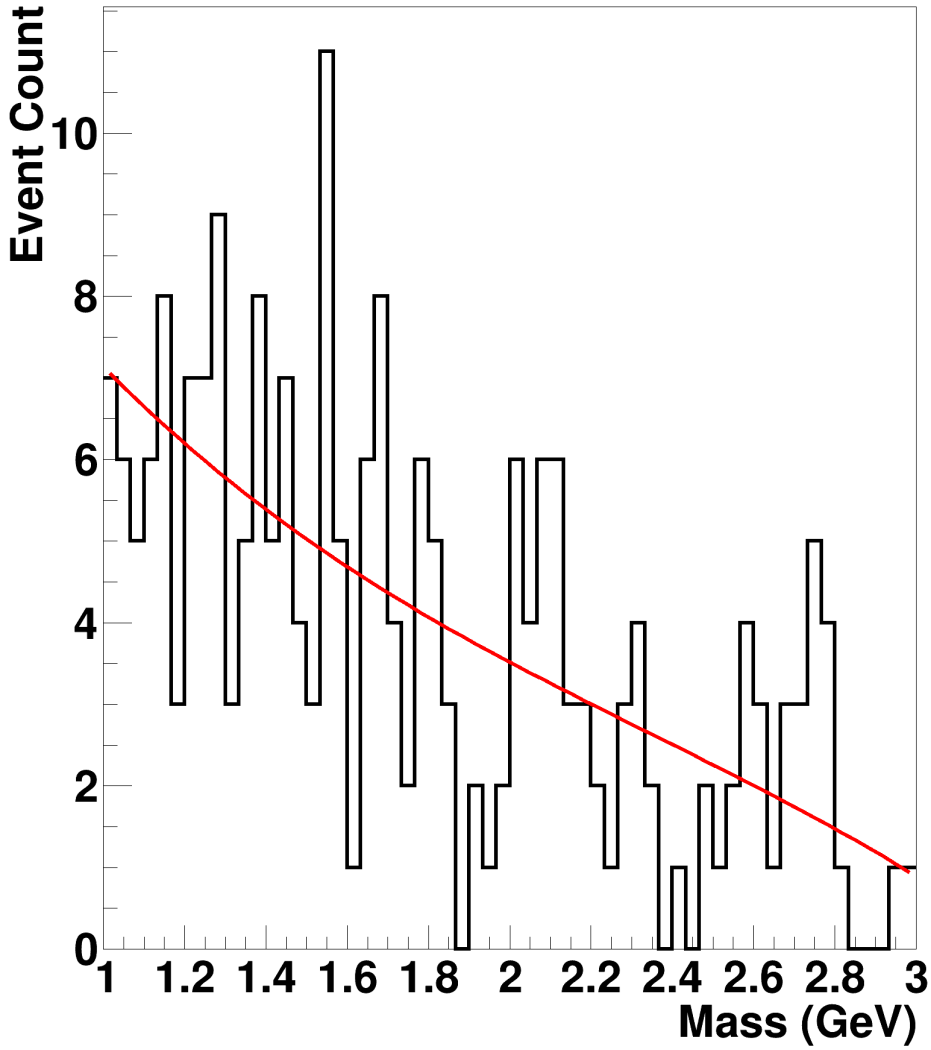

Figure 6: Fit of the linear model $M_{a,b,c,d}$ to the original data for optimal values $\hat{a} = 439, \hat{b} = -312, \hat{c} = 101, \hat{d} = -14$.