Statistical Methods of Data Analysis (physics716)
Final Project: The q^2 spectrum of $B \to D^* l\nu$ decays

Aarathi Parameswaran

March 13, 2023

1 Task 1

1.1 Task 1 (i): Deriving the negative log-likelihood (question 1)

$$\mathcal{L} = \prod_i \mathrm{poisson(n_i^{data}|\lambda_i = n_i^{exp})} = \prod_i \frac{\lambda_i^{n_i^{\text{data}}} \exp(-\lambda_i)}{n_i^{\text{data}}!}$$

$$n_i^{\text{exp}} = N^{\text{sig}} \times h_i^{\text{sig}} + N^{\text{bkg}} \times h_i^{\text{bkg}}$$

$$\mathcal{L} = \prod_i \frac{(n_i^{\text{exp}})^{n_i^{\text{data}}} e^{-n_i^{\text{exp}}}}{n_i^{\text{data}}!}$$

Taking the negative log-likelihood and discarding terms that do not depend on the parameters N^{sig} and N^{bkg}:

$$-\log \mathcal{L} = \sum_i \left[n_i^{\text{data}} \log n_i^{\text{exp}} - n_i^{\text{exp}} - \log n_i^{\text{data}}! \right].$$

The term $-\log n_i^{\text{data}}!$ can be dropped off as a constant. Substituting for n_i^{exp} in the negative log likelihood:

$$-\log \mathcal{L} = \sum_i \left[n_i^{\text{data}} \log \left(N^{\text{sig}} h_i^{\text{sig}} + N^{\text{bkg}} h_i^{\text{bkg}} \right) - \left(N^{\text{sig}} h_i^{\text{sig}} + N^{\text{bkg}} h_i^{\text{bkg}} \right) \right]$$

1.2 Task 1 (ii): Plot of the normalized signal and background

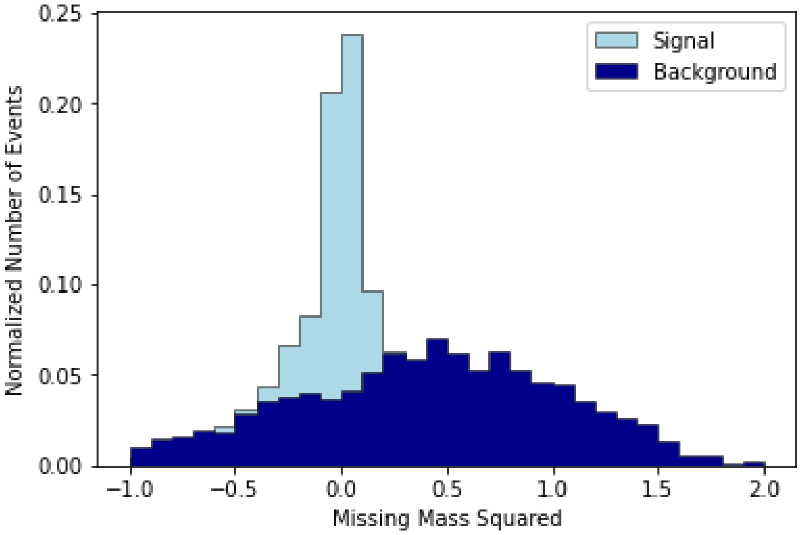

Figure 1: Task 1 (ii): Plots of the normalized signal and background histograms for bin ”q2bin2”

1.3 Task 1 (iii): Fit plot for q^2 bin

Using the derived negative log likelihood and minimizing it using iMinuit, the fit plot is obtained for the same bin.

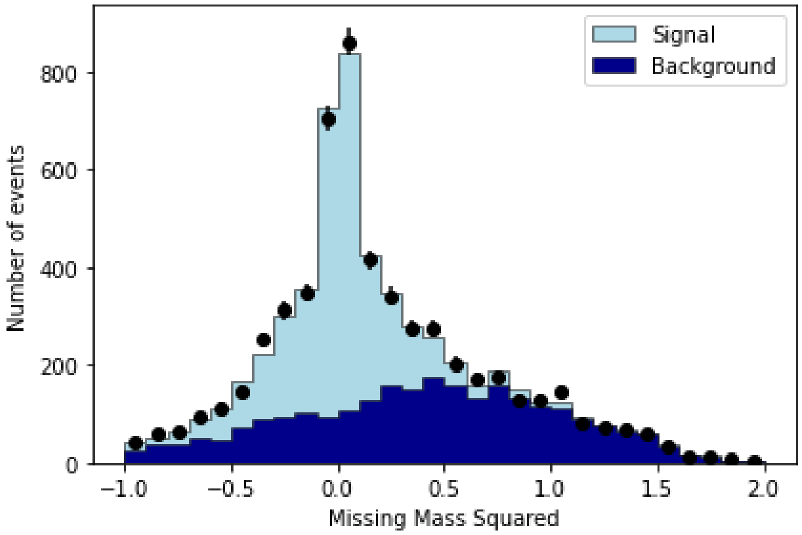

Figure 2: Task 1 (iii): fit plot for the q^2 bin

1.4 Task 1 (iv): χ^2 goodness of fit test

Task 1 (iv) is to conduct a χ^2 goodness of fit test. The obtained χ^2 value is 24.096359401673123. The calculated p value is 0.6764392359842712.

1.5 Questions for task 1

1.5.1 Compute the p-value for the goodness of fit test in Task 1. What does it imply? Also explain your choice for the number of degrees of freedom. (question 2):

Obtained p-value is 0.6764392359842712. This p-value implies that there is a 67.6 percent chance of observing the chi-squared value as larger than that of the one obtained for the null hypothesis being true. If the p-value was lesser than 0.05, we reject the null hypothesis. Since this is not the case, we can assume that this is a good fit to the data.
The number of degrees of freedom is 28, which was obtained by subtracting the two parameters-the background and signal - from the number of observations (which is the number of data points in that bin). They are subtracted because they are parameters estimated from the data.

1.5.2 Discuss methods by which you could further validate the fit results (question 3):

Other methods can also be used to check the goodness of the fit. For example, the Kolmogorov-Smirnov test or using Monte Carlo evaluations. These are commonly used to test given hypotheses.

2 Task 2

2.1 Plot of fitted signal yields and uncertainties in bins of q^2

Task 2 (i) is to perform the fits for all q^2 bins and store the values and errors. Task 2 (ii) involved plotting these fitted signals.

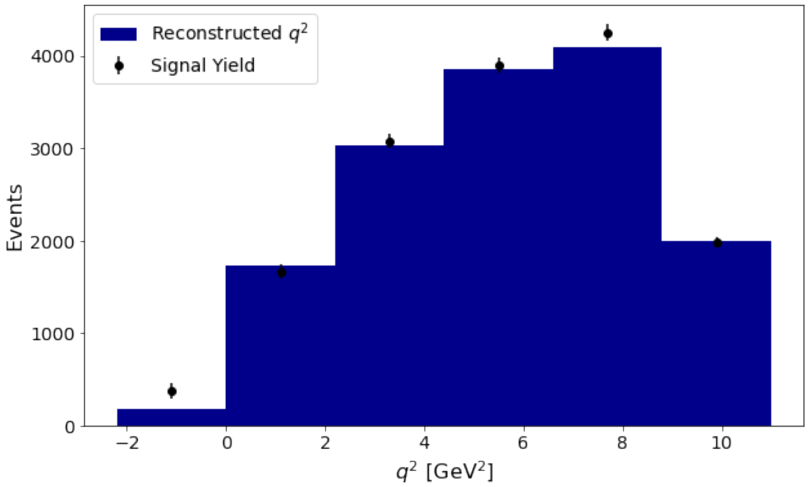

Figure 3: Task 2 (ii): Plot of fitted signal yields and uncertainties with reconstructed q^2 histogram

2.2 Questions

2.2.1 In Task 2 why might there be some disagreement in the negative q^2 bin? Hint: Look explicitly at the fit plot and the fit covariance (question 4)

From the fit plot, it is evident that the discrepancy is largest in the negative q^2 bin. The disagreement in the negative q^2 bin could be because of the number of events in the bin, which is lesser than those of the other q^2 bins, leading to more statistical uncertainties. There could also be systematic uncertainties arising from the detector or the signal. Looking at the covariance matrix, the diagonal elements give the variance, while the off-diagonal elements give the covariance between the parameters. Since the covariance is negative, the parameters are inversely related. This means that when one quantity is overestimated, the other is underestimated. In this case, the background yield is probably overestimated, leading to an underestimation of the signal in this bin.

3 Task 3

3.1 Task 3 (i): Distributions of true and reconstructed q^2 values

The true and reconstructed distributions are plotted. It can be seen that the reconstructed distribution is spread across q^2 more than that of the true values. This could be due to detector effects, where factors like resolution affect the true values.

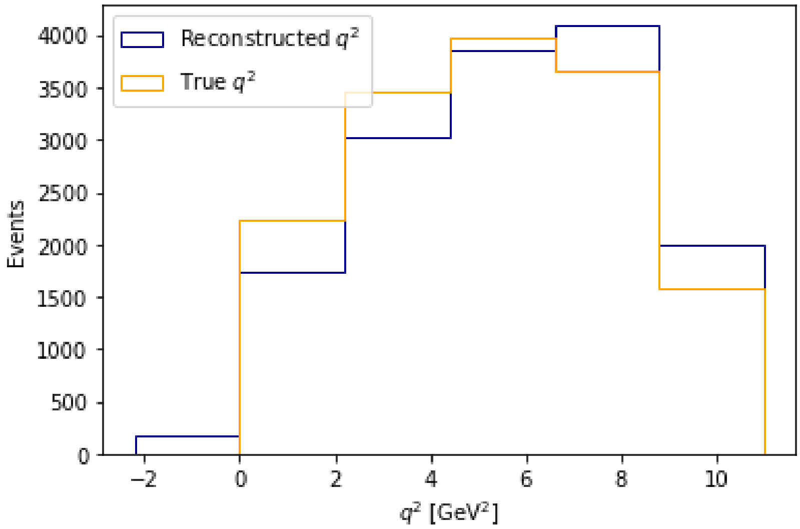

Figure 4: Task 3 (i): True and reconstructed q^2 distributions

3.2 Task 3 (ii): Migration matrix

The migration matrix is defined and visualised, as described in the notebook.

3.3 Task 3 (iii): Unfolding

The aim of unfolding is to obtain the true values from the reconstructed values. This is done by using minuit, and the unfolded values are extracted from this output.

3.4 Task 3 (iv): Plotting unfolded spectrum

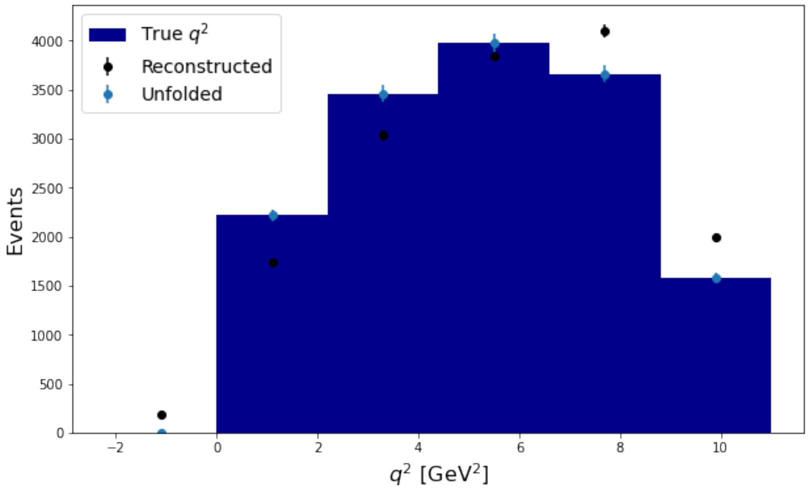

Figure 5: Task 3 (iv): Plot of reconstructed, unfolded and true q^2 distributions

3.5 Questions

3.5.1 How could you improve the unfolding in Task 3? (question 5)

Unfolding, particularly matrix unfolding is generally bad. Using another method for unfolding can improve/compare the results, such as regularized unfolding using pyunfold.

4 References:

- Von Torne, Eckhard (2022), *Lecture 8: Least squares and χ^2 numerical optimization*, Statistical Methods of Data Analysis, University of Bonn.

- Von Torne, Eckhard (2022), *Lecture 9:Monte Carlo methods*, Statistical Methods of Data Analysis, University of Bonn.

- Von Torne, Eckhard (2023), *Lecture 14: Unfolding*, Statistical Methods of Data Analysis, University of Bonn.