School of Physics and Astronomy

DAH Project
Physics

Project F1: Make Accurate Measurements of Υ Masses

Matthew Stewart
December 2020

Abstract

LHCb data has been analysed to fit muon pair mass data. Maximum likelihood estimation was used within an iPython notebook. The $\Upsilon(1S)$ peak had a resolution of 0.04301(16) GeV/c^2 and a mean of 9.4561(3) GeV/c^2. The $\Upsilon(2S)$ peak had a resolution of 0.0460(5) GeV/c^2 and a mean of 10.0192(6) GeV/c^2. The $\Upsilon(3S)$ peak had a resolution of 0.0466(16) GeV/c^2 and a mean of 10.3510(13) GeV/c^2.

Declaration

I declare that this project and report is my own work.

Signature: Date: 04/12/20

Course Organisers: Matt Needham 5 Weeks

Contents

1 Introduction

'Quarkonium hadroproduction' is a modern research topic within QCD looking into hadron production. To test and verify new models, accurate experimental data of meson production is required. At the LHCb experiment, data from pp collisions has been collected on the production of $\Upsilon(1S)$, $\Upsilon(2S)$ and $\Upsilon(3S)$ by The LHCb Collaboration [1]. A plot of the mass spectrum of the muon pair $\mu^+\mu^-$ shows three peaks, corresponding to $\Upsilon(1S)$, $\Upsilon(2S)$ and $\Upsilon(3S)$ shown in figure 1. Each had a mean and resolution which was be estimated using a 'maximum likelihood' fit.

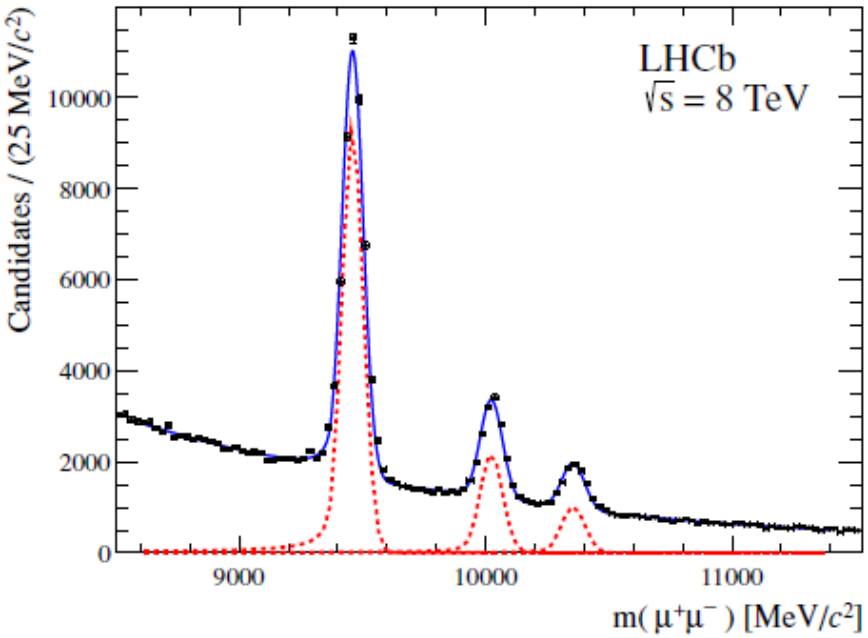

Figure 1: The muon pair mass apectrum obtained by the The LHCb Collaboration [1] showing three peaks corresponding to each of the production of Υ mesons.

2 Theory

2.1 Maximum Likelihood Estimation

To model data with a probability density function (PDF) requires a best-estimate of the parameters needed to compute the PDF. Consider sample data x with individual points x_i. For a general PDF with a set of parameters, τ, of some form. There is a probability density $P(x_i \mid \tau)$ - conditional probability density of obtaining the data point given the set of PDF parameters. The likelihood L, joint probability of obtaining the full sample x given τ, is the product of P for each x_i - shown in equation 1.

$$L = \prod_i P(x_i \mid \tau) \tag{1}$$

The sample data x from experiment is fixed and assumed to be the best description

of the underlying parent distribution. Therefore, the best estimates of τ will maximise L. Consider 'negative log likelihood' (NLL), $-log(L)$, given by equation 2.

$$NLL = -log(L) = -\sum_i log(P(x_i \mid \tau)) \tag{2}$$

By the same logic, now the best estimates of τ will minimise NLL. The is important to note as taking a sum instead of a product is more numerically stable [2] and typical optimising functions are minimisers. It is noted that to numerically minimise NLL an 'initial guess' of the parameters is typically required.

2.2 Gaussian Mass Model

Each of the three mass signal peaks were modelled as a gaussian peak on top of background modelled as an exponential distribution. The resulting composite PDF for a slice of data between an upper limit, x_u, and lower limit, x_l - isolating a single peak is given by equation 3.

$$P(x) = \frac{a}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} + \frac{-\lambda(1-a)}{e^{-\lambda x_u} - e^{-\lambda x_l}} e^{-\lambda x} \tag{3}$$

The parameters to be estimated were; σ, the resolution of the gaussian; μ, the mean of the gaussian; λ, the rate parameter of the exponential; a, a parameter between 0 and 1 scaling the relative strength of the signal to the background. It must be noted that this peak must be normalised over the range of x_l to x_u. The gaussian peak should drop to approximately zero at x_l and x_u assuming the whole peak is captured in the slice of x which means the analytical normalisation over $x_l = -\infty$ to $x_u = \infty$ - the prefactor of $\frac{1}{\sigma\sqrt{2\pi}}$ to the gaussian term. The exponential must also be normalised 'by hand' - the prefactor of $frac-\lambda(1-a)e^{-\lambda x_u} - e^{-\lambda x_l}$ to the exponential term.

To fit a model to n peaks on top of the background gives the total composite PDF of equation 4.

$$P(x) = \sum_{j=1}^{n} \left(\frac{a_j}{\sigma_j\sqrt{2\pi}} e^{-(x-\mu_j)^2/2\sigma_j^2}\right) + \frac{-\lambda(1-\sum_{j=1}^{n} a_j)}{e^{-\lambda x_u} - e^{-\lambda x_l}} e^{-\lambda x} \tag{4}$$

Now each peak has its own σ, μ, and a. The PDF is still normalised over the entire sample data range containing the n peaks.

2.3 Sideband Subtraction

To estimate the number of background eents and signal events in a peak region, the method of 'sideband subtransaction' can be used. A signal region under the peak and two equal width bands equidistant from the peak position on either side are defined as in figure 2 from a separate LHCb experiment [2] included only as an example.

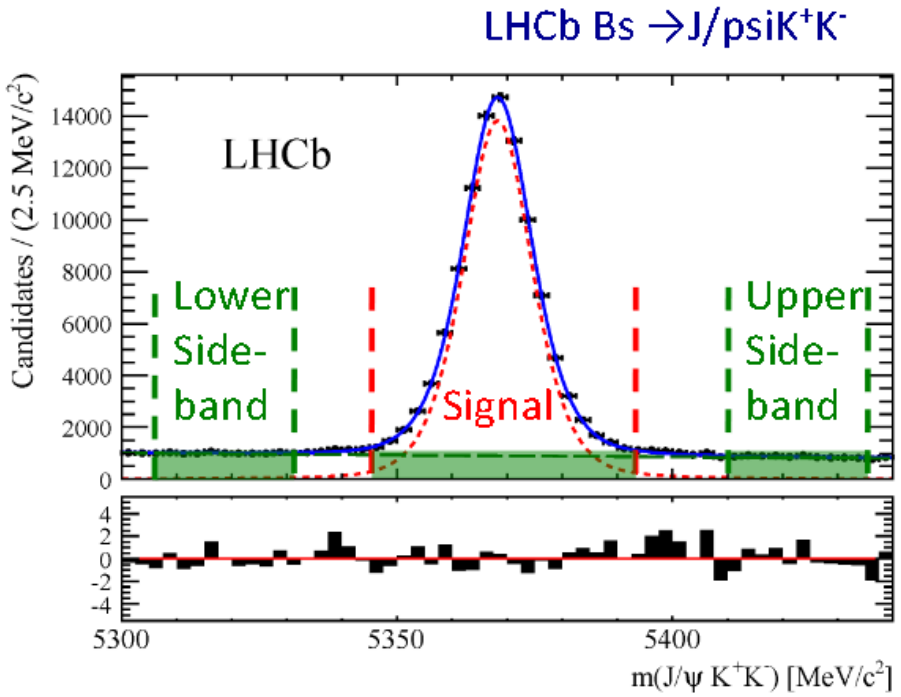

Figure 2: An example of equal width sidebands defined equidistant, either side, from a central peak. Figure taken from coursbooklet [2].

The number of signal events, S, is related to the number of events N, the number of events in the lower sideband, B_{LS} and the number of events in the upper sideband, B_{US}, for sidebands each half the width of the signal region, by equation 5.

$$S = N - (B_{LS} + B_{US}) \tag{5}$$

3 Method

The maximum likelihood fit was carried out in an iPython notebook. The data was imported using as recommended in the coursebooklet. The 'Pandas' library was used to aid in the general understanding of the data initially. Histograms were first plotted for all six of the variables imported. However, all but the muon pair mass spectrum and rapidity looked very compressed at the low end of the x-axis as if the x-axis scale was distorted by outliers. Statistical outliers have a 'z-score', in one or more of the imported variables, outwith the range of -3 to $+3$ [3]. The histograms obtained after the data was cleaned of outliers (9475 points) using stats.zscore() is shown in figure 3.

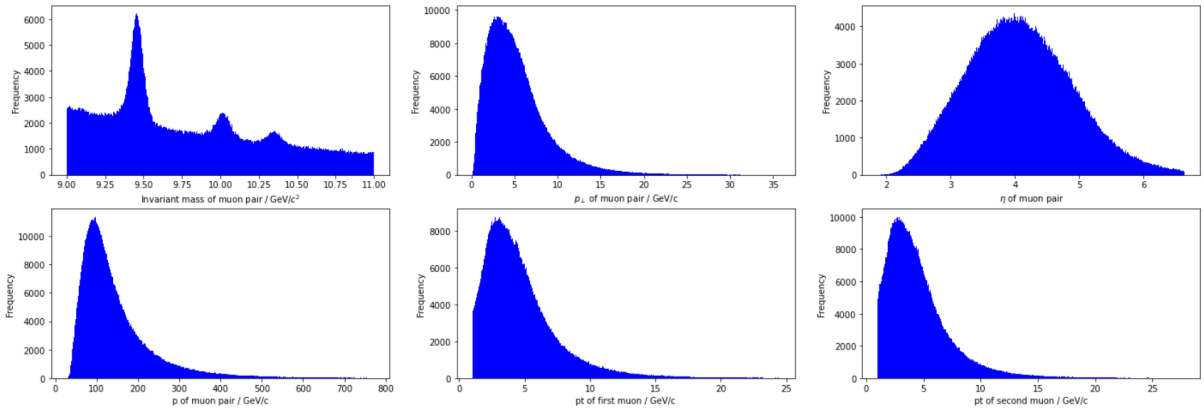

Figure 3: Histograms of the six variables, including the mass spectrum (top left), after statistical outliers had been dropped.

The full muon pair mass spectrum sample was split into three regions, each containing one peak. The three peak positions were estimated by the most populated bins. For the lowest mass peak corresponding to $\Upsilon(1S)$, the full width at half maximum ($FWHM$) was estimated from visual inspection. Assuming a gaussian peak, the resolution σ is related to $FWHM$ by $FWHM \approx 2.35\sigma$. From this expression, σ was estimated for the first peak.

Sideband subtraction was used for a region of $\pm 150 Mev/c^2$ either side of the $\Upsilon(1S)$ peak to estimate the number of signal events compared to the number of background events in the peak region.

A function was written to return the value for the gaussian- exponential composite PDF given a set of parameters outlined in 2.2. A second function returned the negative log-likelihood (NLL) for the split mass data. This function was passed though scipy.optimize.minimize() which iteratively shifted the parameters in the direction minimising the NLL. The minimisation required an initial guess at the parameters which was decided by the previous estimates of σ_1 and μ_1 as well as some trial and error.

This returned a set of best-estimate parameters based on maximum likelihood estimation. By inputting these into the PDF function, the model PDF was plotted on top of a normalised histogram of the data shown in figure 4.

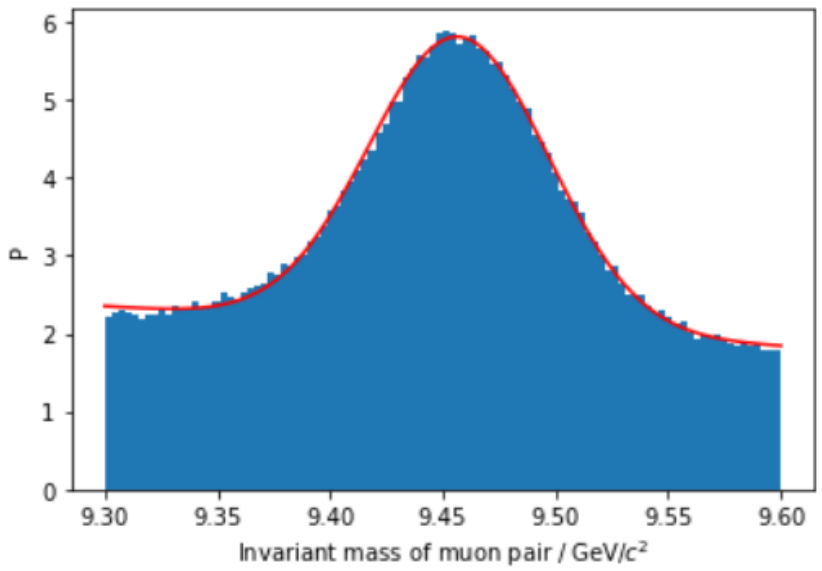

Figure 4: The blue bars are the normalised histogram of the isolated $\Upsilon(1S)$ peak in the mass spectrum. The red curve is the PDF generated by maximum likelihood estimation.

The scipy.optimize.minimize() function had the option to declare a minimisation method. The 'Nelder-Mead' method was most reliable and fastest so was used to verify initial guesses. Once good guesses had been established, the fit was redone with the 'BFGS' method. The advantage of this was that the 'BFGS' method also returned the inverse of the hessian matrix, called the covariance matrix, of which the diagonal elements give the estimates on the square of the parameter errors [4]. This was repeated for the two other mass ranges corresponding to the $\Upsilon(2S)$ and $\Upsilon(3S)$ peaks.

The results of these fits inspired initial guesses at the parameters of a combined gaussian model PDF of the form outlined in equation 4 for $n = 3$ peaks. Best-estimate parameters were found by the same maximum likelihood approach as for the single peak. This model combined PDF was plotted on top of the normalised histogram of the full mass sample data. The residuals - difference between the normalised histogram value and the PDF evaluated at the histogram bin centre - were plotted to investigate the accuracy of the fit.

4 Results

The three-peak, combined PDF determined by maximum likelihood estimation is is plotted on top of the data in figure 5.

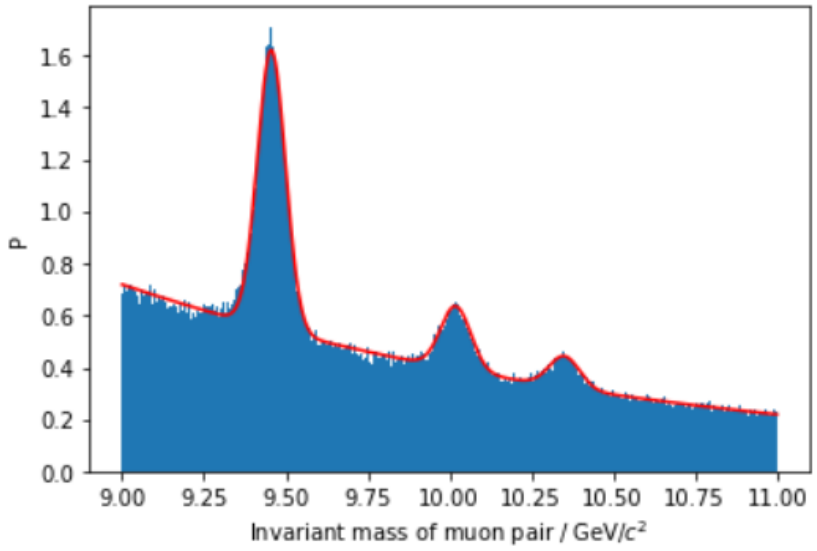

Figure 5: The blue bars are the normalised histogram of the full mass spectrum. The red curve is the PDF generated by maximum likelihood estimation.

The best estimate parameters are presented in table 1.

Parameter	**Best Estimate**	**Error**^{2}	**Error**
σ_1	0.04301	2.75851830e-08	0.00016
μ_1	9.45606411	6.25066090e-08	0.0003
σ_2	0.04595628	2.62163512e-07	0.0005
μ_2	10.01921073	3.50436651e-07	0.0006
σ_3	0.04663122	2.54225360e-06	0.0016
μ_3	10.35097279	1.66988824e-06	0.0013

Table 1: The best estimate parameters returned by the combined fit. The square of the errors were given by the diagonal of the hessian inverse matrix.

Therefore, the $\Upsilon(1S)$ peak had a resolution of 0.04301(16) GeV/c^2 and a mean of 9.4561(3) GeV/c^2. The $\Upsilon(2S)$ peak had a resolution of 0.0460(5) GeV/c^2 and a mean of 10.0192(6) GeV/c^2. The $\Upsilon(3S)$ peak had a resolution of 0.0466(16) GeV/c^2 and a mean of 10.3510(13) GeV/c^2.

5 Discussion

From visual inspection, the fit depicted in figure 5 looks reasonably accurate. However, when carrying out the minimisation of NLL, the 'BFGS' method encountered various Python errors. The curve plotted is the result of a fit using the 'Nelder-Mead' method. This does not return the square of the parameter errors like 'BFGS' so the quoted errors come from the results of the individual fits of the peaks and not the combined PDF of all three together. Even then, there is 'precision loss' and error precision cannot be guaranteed. These fits had fewer free parameters so were likely more precise than the combined fit of all three so these errors are likely underestimated. The cause of the Python errors

is unknown but other minimisers could have been tried such as scipy.optimize.curvefit() for example.

The plot of the residuals in shown in figure 6.

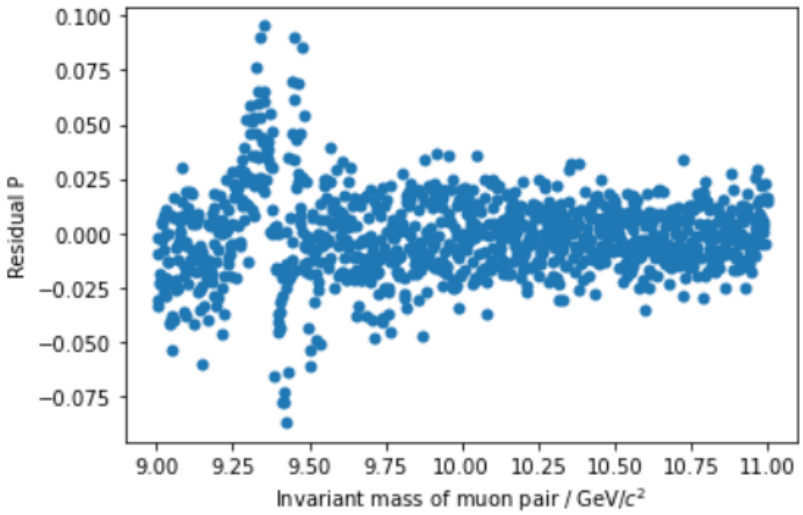

Figure 6: The points are the residuals of P, the normalised histogram of the sample data minus the model prediction.

From inspection, the residual data points are randomly spread above and below the x-axis so the best estimates are likely accurate. For masses around the first peak ≈ 9.45 GeV/c^2, there is the most deviation of the model from the sample data. This should be reflected in the errors on σ_1 and μ_1 being more significant than the others but that is not found by the results likely due to the issues in estimating the errors for the combined PDF. The sharpness of the first peak means that the actual sample data forming it changes a lot from one bin to the next compared to the other peaks. This less 'smoothness' is expected to be more difficult to fit. This means it is not surprising that a weakness in the fit could be located at ≈ 9.45 GeV/c^2.

A more advanced model would be to use a sum of 'crystal-ball' functions instead of gaussian peaks [1]. This improved method should produce parameters with smaller relative errors than the gaussian mass model.

6 Conclusion

LHCb data from pp collisions has been analysed to estimate parameters in the fitting of the muon pair mass spectrum. Maximum likelihood estimation was used within an iPython notebook. The mean and resolution of three peaks in the muon pair mass spectrum have been estimated using maximum likelihood estimation. The $\boldsymbol{\Upsilon}(1S)$ peak had a resolution of 0.04301(16) GeV/c^2 and a mean of 9.4561(3) GeV/c^2. The $\boldsymbol{\Upsilon}(2S)$ peak had a resolution of 0.0460(5) GeV/c^2 and a mean of 10.0192(6) GeV/c^2. The $\boldsymbol{\Upsilon}(3S)$ peak had a resolution of 0.0466(16) GeV/c^2 and a mean of 10.3510(13) GeV/c^2. The mean values of these estimates are accurate but the values of the errors are likely inaccurate do to issues with the code.

7 References

References

[1] R Aaij, C Abellan Beteta, B Adeva, M Adinolfi, C Adrover, A Affolder, Z Ajaltouni, J Albrecht, F Alessio, M Alexander, S Ali, G Alkhazov, P Alvarez Cartelle, A A Alves, S Amato, S Amerio, Y Amhis, L Anderlini, J Anderson, and R Andreassen. Production of J/ψ and Υ mesons in pp collisions at $\sqrt{s} = 8$ TeV. *Journal of high energy physics : JHEP.*, 2013(6), 2013. 2, 8

[2] Matt Needham. Data acquisition and handling 2019-20 coursebooklet, September 2020. 3, 4

[3] Erin Dienes. I Have An Outlier! . `https://www.ctspedia.org/do/view/CTSpedia/OutLier`, 2011. 4

[4] Wikipedia. Covariance Matrix. `https://en.wikipedia.org/wiki/Covariance_matrix`, 2020. 6