S1: Principals of Data Science - Coursework

William Knottenbelt, wdk24

December 9, 2023

Section A

Part (a)

The total probability density function, p, is properly normalised over $M \in [-\infty, +\infty]$ if it integrates to unity. The integral of p can be written as a weighted sum of signal and background terms:

$$\int_{-\infty}^{+\infty} p\, dM = f \int_{-\infty}^{+\infty} s\, dM + (1-f) \int_{0}^{+\infty} b\, dM. \tag{1}$$

To show that p is properly normalised we require the identity:

$$\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}, \tag{2}$$

For the signal term, we have:

$$\int_{-\infty}^{+\infty} s\, dM = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) dM = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx,$$

where we used the substitution $x = M - \mu$. This integral has the form of identity (2) with $a = \frac{1}{2\sigma^2}$, hence we see that:

$$\int_{-\infty}^{+\infty} s\, dM = \frac{1}{\sigma\sqrt{2\pi}} \sqrt{2\sigma^2\pi} = 1.$$

For the background term we have:

$$\int_{0}^{+\infty} b\, dM = \int_{0}^{+\infty} \lambda e^{-\lambda M}\, dM = (-e^{-\lambda M})|_{0}^{+\infty} = 1$$

Hence, both the signal and background terms integrate to unity. We can plug this into equation 1 and we see that the normalisation condition is satisfied:

$$\int_{-\infty}^{+\infty} p\, dM = f + (1-f) = 1.$$

Part (b)

In general, any probability density function, $g(X)$, can be normalised over an arbitrary range $[\alpha, \beta]$ by multiplying it by the factor $\frac{1}{F(\beta)-F(\alpha)}$, where F is its respective cumulative density function, CDF. To see proof of this, we add normalisation factor, A, to g. We have:

$$1 = \int_\alpha^\beta Ag(M)\,dM = A\left(\int_{-\infty}^\beta g(M)\,dM - \int_{-\infty}^\alpha g(M)\,dM\right) = A(F(\beta) - F(\alpha))$$

Hence:

$$A = \frac{1}{F(\beta) - F(\alpha)}$$

To ensure that the signal and background distributions contribute the correct fractions $(f, (1-f))$ to the total probability, they must be normalised individually before summing them. The CDF of the normal distribution is $F(X) = \Phi(\frac{X-\mu}{\sigma})$, hence the normalisation factor for the signal distribution is:

$$\frac{1}{\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})}$$

The CDF of the exponential decay distribution is:

$$F(X) = \begin{cases} 1 - e^{-\lambda X} & \text{for } X \geq 0, \\ 0 & \text{for } X < 0. \end{cases}$$

Hence, provided that $\alpha, \beta > 0$, the normalisation factor for the background distribution is:

$$\frac{1}{e^{-\lambda\alpha} - e^{-\lambda\beta}}.$$

The total probability density function is a weighted sum of the signal and background distributions. Since the signal and background are both normalised, the total PDF is also normalised:

$$\int_\alpha^\beta p\,dM = f\int_\alpha^\beta s\,dM + (1-f)\int_\alpha^\beta b\,dM = f + (1-f) = 1.$$

Assuming $\alpha, \beta > 0$, the full expression is:

$$p(M) = \frac{f}{\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) + (1-f)\frac{\lambda e^{-\lambda M}}{e^{-\lambda\alpha} - e^{-\lambda\beta}}.$$

Part (c)

We implemented the normalised signal, background and total PDFs as found in part (b) using the `scipy.stats` package. We then integrated the total PDF over $[5, 5.6]$ for 1000 random combinations of the parameters, $\boldsymbol{\theta}$, and all integrals came out to unity as expected.

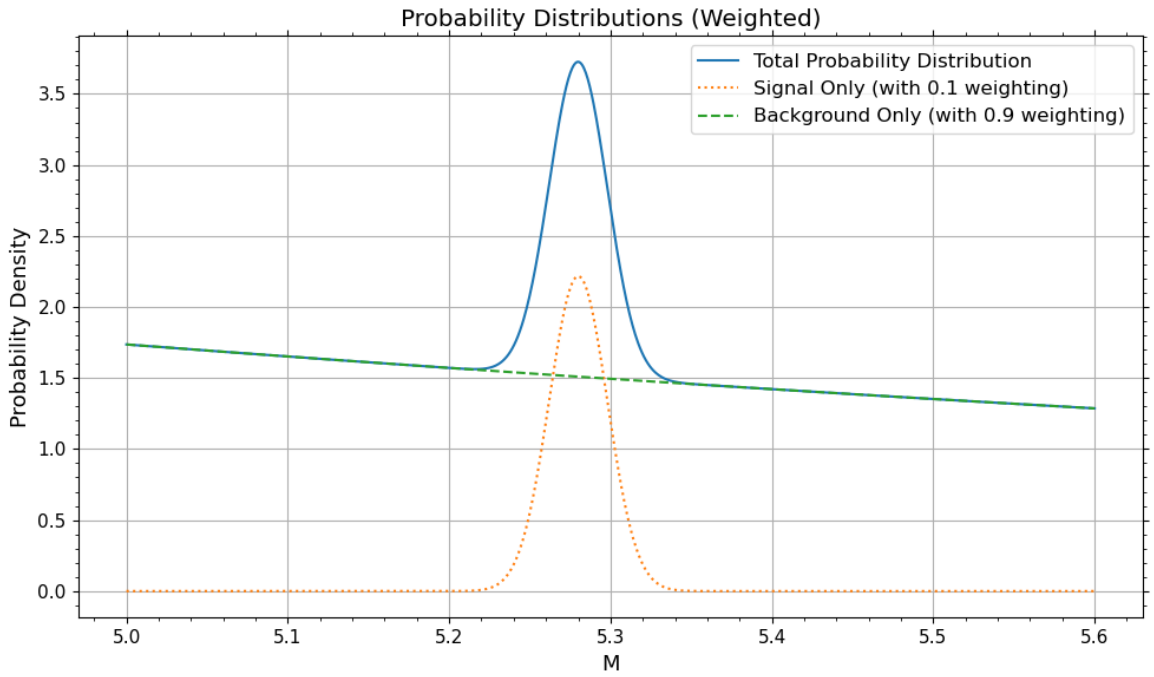

Figure 1: Visualisation of the total, signal and background PDFs with the weights applied to the signal and background models.

Part (d)

We created two plots visualising the distributions with the true parameters. Fig. 1 includes the weights f and $(1 - f)$ on the signal and background distributions. Fig. 2 shows the distributions properly normalised (ie. without the weights).

Part (e)

To generate samples from the total PDF, we used the inverse CDF method, which works by generating uniform random numbers in [0, 1] and passing them into the inverse CDF (A.K.A percentage point function, PPF) to generate events distributed by the original PDF. Since we did not have access to the PPF of the total distribution, we implemented an algorithm which, for each event, chooses whether to generate from the signal-only or background-only distributions with probability 0.1 and 0.9, then generates the event using the PPF of the chosen model. This method is significantly more efficient than the accept-reject method, since there are no wasted samples. We used the PPFs of the normal and exponential distributions available in `scipy.stats`, which are not automatically normalised over $[\alpha, \beta]$, hence rather than generating uniform random numbers in the interval [0, 1], we generated from the interval $[F(\alpha), F(\beta)]$, where F is the CDF of the distribution we are generating from. This guarantees that data will only be generated in the desired range, and it will be distributed by the correctly normalised PDF (since only the relative probability of two points in the range $[\alpha, \beta]$ matters, which the same for a non-normalised model).

We generated a sample of 100K events, then fitted the total PDF to this data using maximum

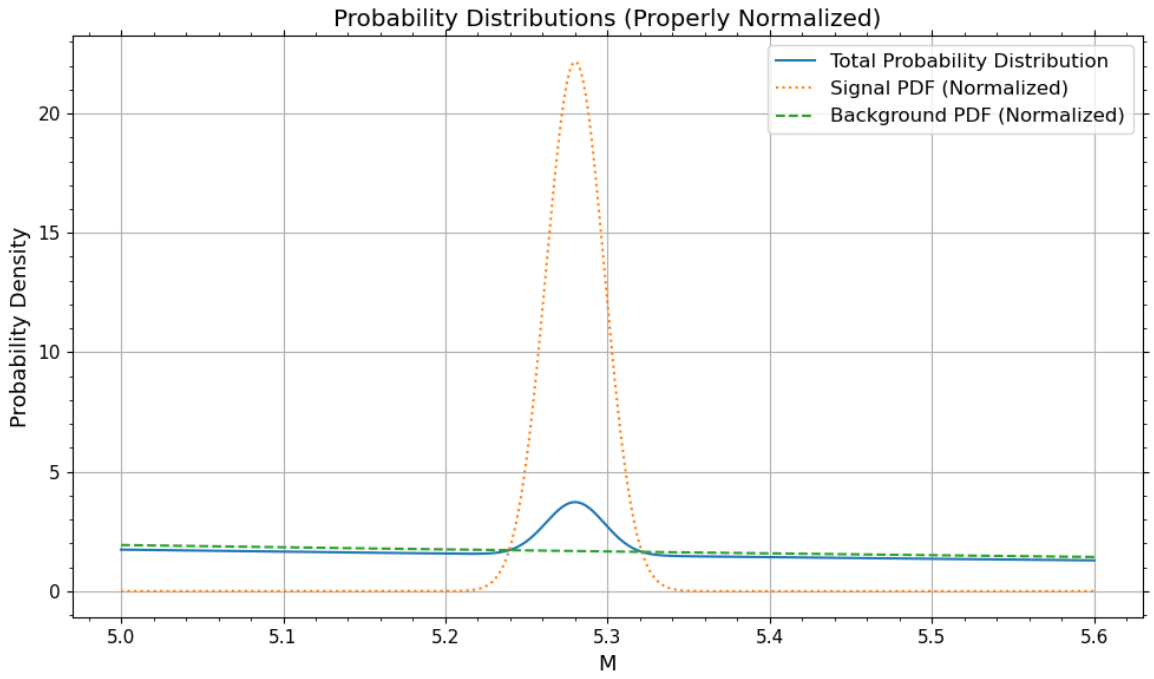

Figure 2: Visualisation of the total, signal and background PDFs, with each distribution properly normalised.

likelihood estimation of the parameters. That is, we estimated the parameters by minimising the negative log likelihood:

$$l = -ln\left(\prod_i p(M_i;\boldsymbol{\theta})\right) = -\sum_i ln(p(M_i;\boldsymbol{\theta})) \tag{3}$$

For a given maximum likelihood estimate $\hat{\theta} \in \{\hat{f}, \hat{\lambda}, \hat{\mu}, \hat{\sigma}\}$, the variance of that estimate is given by the minimum variance bound (AKA the Cramér-Rao lower bound) in the asymptotic limit of large sample sizes ($N \to \infty$):

$$V(\hat{\theta}) = \left(NE\left[\frac{\partial^2 l}{\partial \theta^2}\right]\right)^{-1}, \tag{4}$$

where n is the sample size, $E[\cdot]$ is the expectation operator, l is the negative log likelihood. A useful property of maximum likelihood estimation is that the expectation of the double differential of the log likelihood is equal to the evaluation of the double differential at the estimated value. Hence:

$$V(\hat{\theta}) = \left(N\left[\frac{\partial^2 l}{\partial \theta^2}\right]_{\theta=\hat{\theta}}\right)^{-1} \tag{5}$$

Since we are dealing with a sample of 100K events, this asymptotic limit is a reasonable assumption. In this limit, the maximum likelihood estimate produces an estiamte which is normally distributed, unbiased, consistent and maximally efficient (minimum variance as discussed). This estimation was

done using `iminuit`, using starting parameters as the true parameters shifted by an appropriate random shift. We obtained the following estimates:

$$\hat{f} = 0.0998 \pm 0.0016, \quad \hat{\lambda} = 0.470 \pm 0.019, \quad \hat{\mu} = 5.27980 \pm 0.00033, \quad \hat{\sigma} = 17.99 \times 10^{-3} \pm 0.32 \times 10^{-3}.$$

For all estimates, the true values of the parameters lie within the uncertainties of the estimates. We then binned and plotted the sample, and overlaid the fitted PDF, as can be seen in Fig. 3.

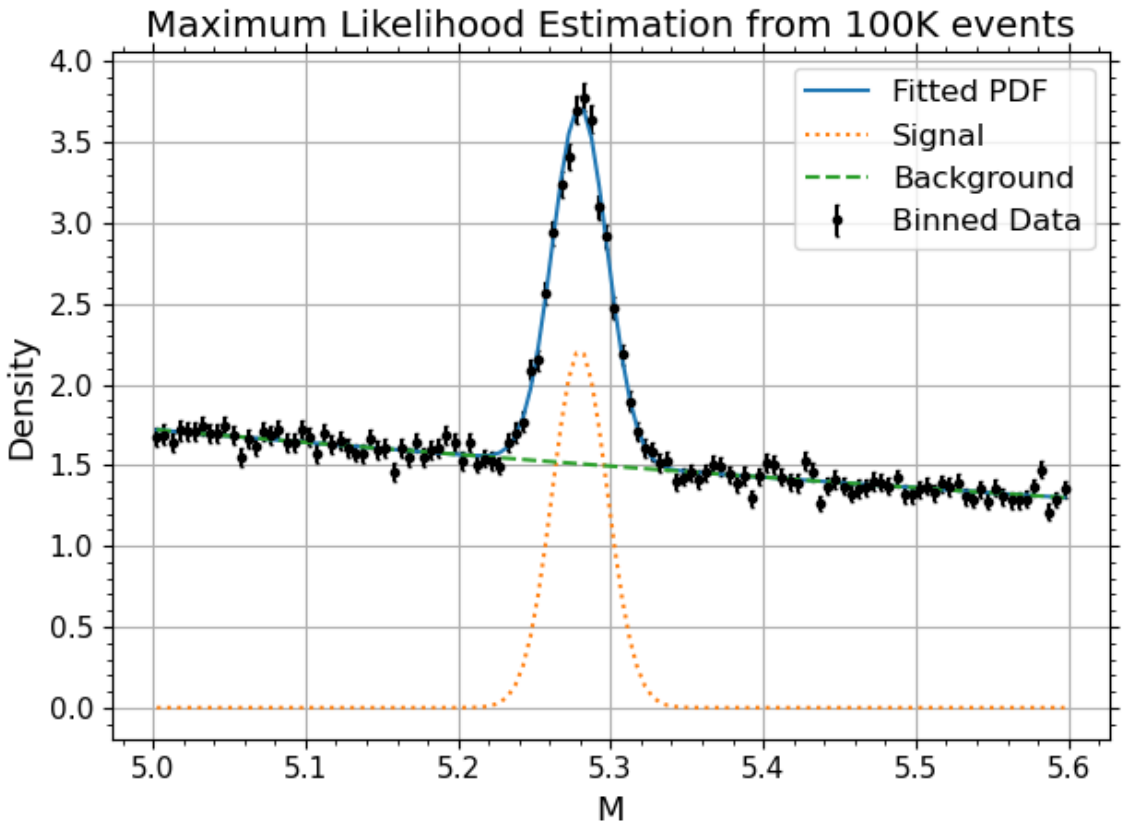

Figure 3: Visualisation of the our 100K sample generated from the true PDF, overlaid by the fitted PDFs of the total, signal and background distributions, whose parameters were estimated via the maximum likelihood method