S1: Principals of Data Science - Coursework

William Knottenbelt, wdk24

December 7, 2023

Part (a)

The total probability density function is given by:

$$p(M; f, \lambda, \mu, \sigma) = fs(M; \mu, \sigma) + (1 - f)b(M; \lambda),$$

where $s(M; \mu, \sigma)$ is the normal distribution and $b(M; \lambda)$ is the exponential decay distribution, which is only non-zero for $M \geq 0$.

The condition for $p(M; f, \lambda, \mu, \sigma)$ to be properly normalised over $M \in [-\infty, +\infty]$ is:

$$\int_{-\infty}^{+\infty} p \, dM = 1,$$

To show that p is properly normalised we will use the identity:

$$\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}, \tag{1}$$

We have:

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

For the signal term, we have:

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

We use the substitution $x = M - \mu$ such that $dx = dM$ and the limits are the same since $x(M \to \pm\infty) \to \pm\infty$ for any finite μ. Then we have:

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

Using identity (**??**) with $a = \frac{1}{2\sigma^2}$, we find that:

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

For the background term we have:

$$\int_0^c b\,dM = \int_0^c \lambda e^{-\lambda M}\,dM = (-e^{-\lambda M})|_0^c = 1 - e^{-\lambda c} \xrightarrow[c\to+\infty]{} 1.$$

Hence, we see that:

$$\int_{-\infty}^{+\infty} p\,dM = f\int_{-\infty}^{+\infty} s\,dM + (1-f)\int_0^{+\infty} b\,dM = f + (1-f) = 1.$$

Thus, the normalisation condition is satisfied for p over $M \in [-\infty, +\infty]$.

Part (b)

Throughout this report, we assume that f being the 'fraction of the signal' means that f is the fraction of the total probability that the signal distribution contributes to the combined probability distribution. To ensure that the signal distribution contributes the correct fraction f to the total probability, and the normal distribution contributes $(1-f)$, we must first normalise each distribution individually over the range $M \in [\alpha, \beta]$, before performing a weighted sum to construct the total PDF. As each distribution is normalised separately, it is guaranteed that any weighted sum of the distributions will also be correctly normalised (provided that the weights sum to unity).

For M restricted to the range $M \in [\alpha, \beta]$, the signal distribution is defined:

$$s(M;\mu,\sigma) = \begin{cases} \frac{A}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) & \text{for } \alpha < M \leq \beta \\ 0 & \text{otherwise.} \end{cases}$$

Where A is a normalisation factor, μ is the mean of the distribution and σ is the standard deviation. For s to be properly normalised we must have:

$$\int_\alpha^\beta s(M)\,dM = 1.$$

Hence:

$$\begin{aligned} 1 = \int_\alpha^\beta s(M)\,dM &= \int_{-\infty}^\beta s(M)\,dM - \int_{-\infty}^\alpha s(M)\,dM \\ = A\left(\int_{-\infty}^\beta \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right)dM - \int_{-\infty}^\alpha \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right)dM\right) \\ &= A(F_{norm}(\beta) - F_{norm}(\alpha)), \end{aligned}$$

where F_{norm} is the cumulative density function of the normal distribution defined over the range $[-\infty, +\infty]$, which is given by:

$$F_{norm}(X) = \Phi(\frac{X-\mu}{\sigma}).$$

Thus, the normalisation factor is:

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

Assuming that $\alpha, \beta > 0$, the background distribution is defined:

$$b(M;\lambda) = \begin{cases} B\lambda e^{-\lambda M} & \text{for } \alpha < M \leq \beta, \\ 0 & \text{otherwise,} \end{cases}$$

where B is a normalisation factor and λ is the decay constant. For b to be properly normalised, we must have:

$$\begin{aligned} 1 = \int_{\alpha}^{\beta} b(M)\, dM &= \int_{-\infty}^{\beta} b(M)\, dM - \int_{-\infty}^{\alpha} b(M)\, dM \\ &= B\left(\int_{0}^{\beta} \lambda e^{-\lambda M}\, dM - \int_{0}^{\alpha} \lambda e^{-\lambda M}\, dM\right) \\ &= B(F_{exp}(\beta) - F_{exp}(\alpha)), \end{aligned}$$

where F_{exp} is the cumulative density function of the exponential decay distribution, which is given by:

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

Hence we have:

$$B = \frac{1}{F_{exp}(\beta) - F_{exp}(\alpha)} = \frac{1}{e^{-\lambda\alpha} - e^{-\lambda\beta}}.$$

Finally, the total probability density function, is a weighted sum of the individually normalised distributions:

$$p(M) = fs(M;\mu,\sigma) + (1-f)b(M;\lambda).$$

This is guaranteed to be properly normalised since:

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

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

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

Part (c)

We coded the normalised signal, background and total distributions using the `scipy.stats` package to implement the expressions found in part (b). Namely:

$$s(M;\mu,\sigma) = \frac{1}{\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right),$$

$$b(M;\lambda) = \frac{\lambda e^{-\lambda M}}{e^{-\lambda\alpha} - e^{-\lambda\beta}},$$

$$p(M;\boldsymbol{\theta}) = fs(M;\mu,\sigma) + (1-f)b(M;\lambda),$$

where $\boldsymbol{\theta} = (f, \lambda, \mu, \sigma)$ represents the parameters.

Once I had coded the probability density functions, I wrote a script `solve_part_c.py`, which integrates the total probability density over $M \in [5, 5.6]$ using `scipy.integrate.quad` for 1000 random combinations of the parameters $(f, \lambda, \mu, \sigma)$, and checks that each integral comes out to unity. This script was run and all checks passed successfully.

Part (d)

We created two plots visualising the distributions with the true parameters $f = 0.1, \lambda = 0.5, \mu = 5.28, \sigma = 0.018$. Fig. 1 shows the distributions with the fractions f and $(1-f)$ applied to the signal and background distributions. Fig. 2 shows the distributions properly normalised.

To re-generate these plots, run `solve_part_d.py` and find the results in `plots/`

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 likelihood estimation of the parameters. This was done by minimising the negative log likelihood with

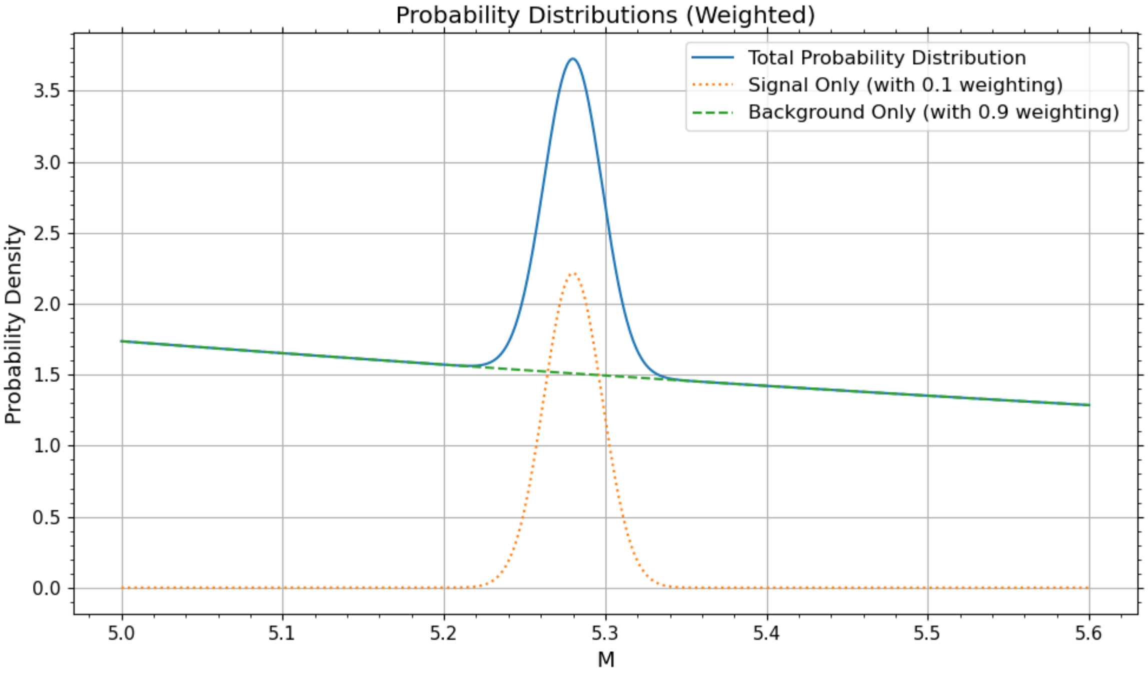

Figure 1: Visualisation of the total, signal and background probability density functions with the weights applied to the signal and background models such that only the total PDF is properly normalised.

`iminuit`, using starting parameters as the true parameters shifted by some random shift. We obtained uncertainties for these estimates from the minimum variance bound (Cramér-Rao lower bound), which is is given by

$$V(\hat{\theta}) = \left(nE\left[\left(\frac{\partial}{\partial\theta} \ln p(M;\boldsymbol{\theta}) \right)^2 \right]_{\theta=\hat{\theta}} \right)^{-1} = \left(nE\left[\left(\frac{\partial^2}{\partial\theta^2} \ln p(M;\boldsymbol{\theta}) \right) \right]_{\theta=\hat{\theta}} \right)^{-1}$$

where n is the sample size, $E[\cdot]$ is the expectation operator, p is the total probability density function

It can be shown that the uncertainty for maximum likelihood estimates is given by the minimum variance bound in the asymptotic limit, which is a reasonable assumption as we are dealing with a large sample of 100K events. 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.

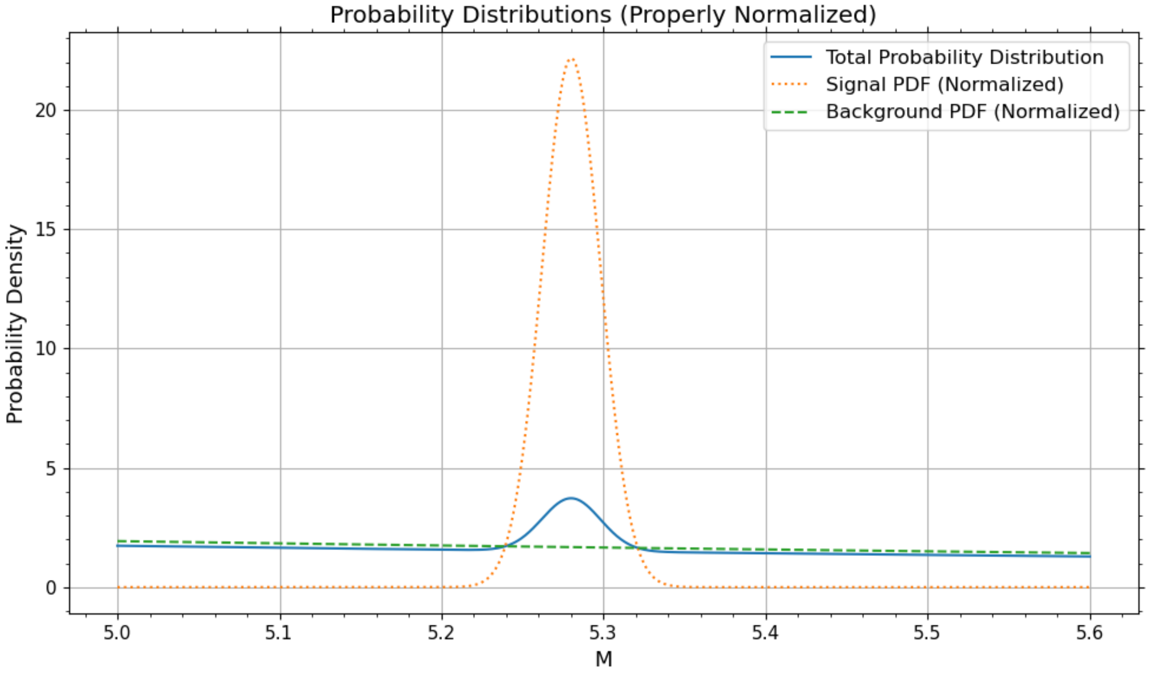

Figure 2: Visualisation of the total, signal and background probability density functions where each distribution is properly normalised.

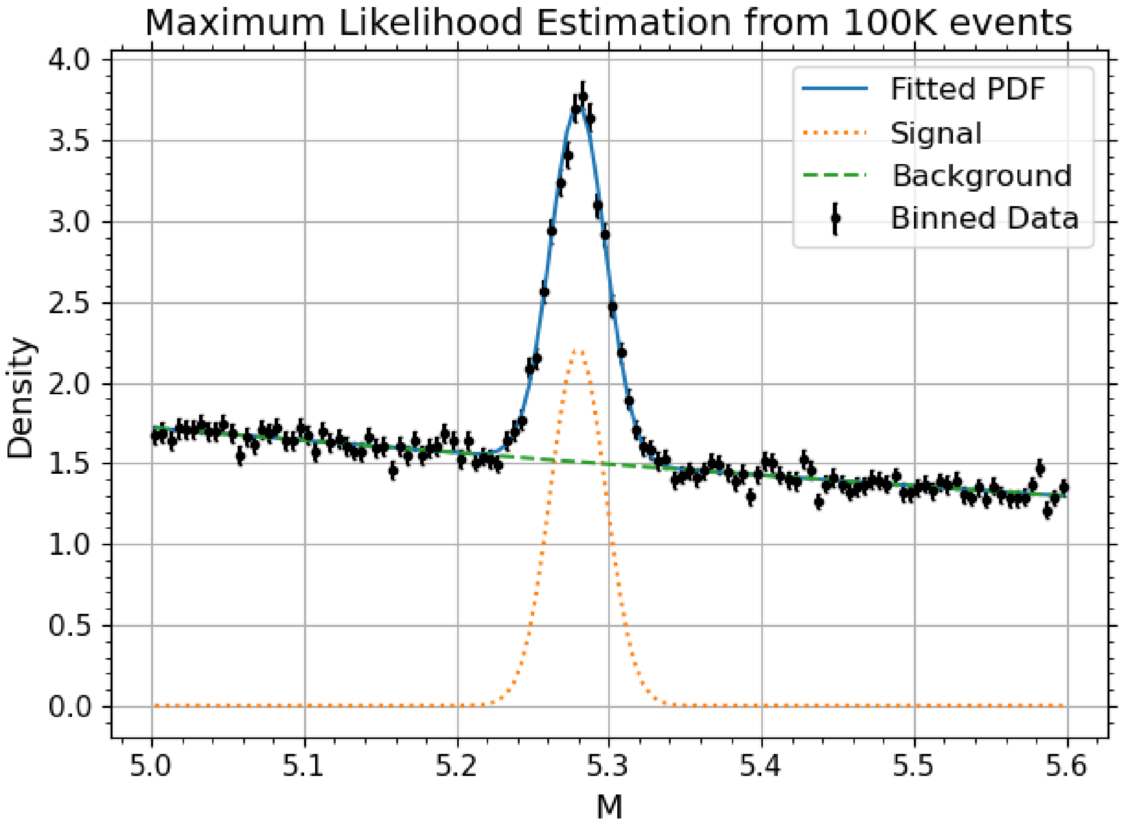

Figure 3: Visualisation of the our 100K sample generated from the true PDF, overlaid by the fitted PDFs of the total, signal and background distributions, whose parameters were estimated via the maximum likelihood method