Exam 2019
Advanced Methods in Applied Statistics

Daniel Ramyar
NJT478

April 5, 2019

I Daniel Ramyar expressly vow to uphold my scientific and academic integrity by working individually on this exam and soliciting no direct external help or assistance.

Problem 1

To solve this problem I start by plotting the data in a histogram and deduce which distributions they most likely are from by looking at the shape of the binned date.

For the data in column 1 it looks like we have some exponentially decaying function which oscillate. The only given function which as both an exponential and a sine is

$$f_8 = sin(ax) + ce^{bx} + 1. \tag{1}$$

For the data in column 2 it looks like we have some second degree polynomial. The only given function which looks like a second degree polynomial is

$$f_6 = 1 + ax + bx^2. \tag{2}$$

For the data in column 3 it looks like it either could come from a binomial or poisson distribution since the data is discrete.

I will have a significance level $\alpha = 0.05$ where the corresponding threeshold χ^2 value will be calculated with the correct degrees of freedom for each fit.

The fit values for the data in column 1 is summarised in figure 1. I get a χ^2 value of 73.40 which is below my threeshold of value 87.10 calculated with 67 degrees of freedom, therefore i cannot reject that the values don't come from (1). And by looking at the fit by eye it looks pretty good, we have to notice though that the fit parameters was very sensitive to the start values so a rasterscan for the likelyhood in the parameter landscape would probably be a good idea to ensure that my fittet parameters are not stuck in local minima.

The fit values for the data in column 2 is summarised in figure 2. I get a χ^2 value of 48.63 which is below my threeshold of value 53.38 calculated with 38 degrees of freedom, therefore i cannot reject that the values don't come from (2). And by looking at the fit by eye it looks pretty good.

The fit values for the data in column 3 is summarised in figure 3. I get a χ^2 value of 15.00 which is below my threeshold of value 32.67 calculated with 21 degrees of freedom, therefore i cannot reject that the values don't come from a poisson distribution. And by looking at the fit by eye it looks pretty good. We note however we get almost the same maximum likelihood value for the binomial distribution for $n = 5000$ and $p = 0.00183219$ which makes sense since the poisson distribution is an approximation of the binomial distribution so both should be able to fit the data.

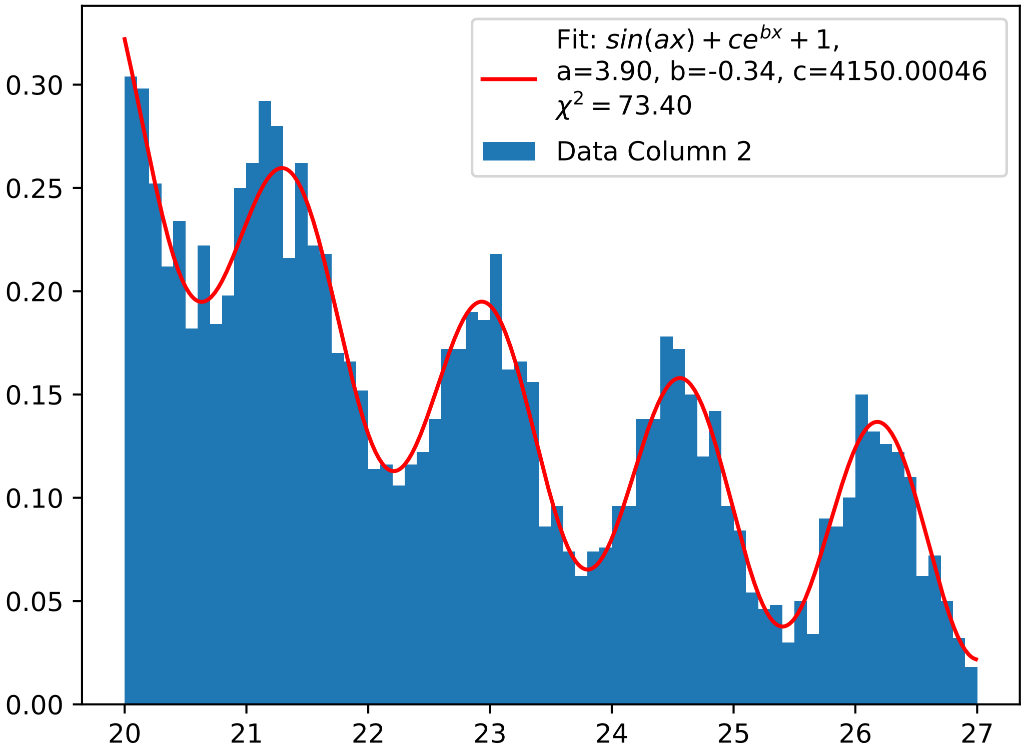

Figure 1: Here we see data from column 1 plotted in a histogram together which its corresponding fit

Problem 2

a)

To begin with i will generate points from a uniform distribution between 0 and 2π for the azimuth angle and from -1 to 1 for the zenith angle.

The generated and given data are shown in figure 4 and 5 where the histogram are normed.

Choosing a significance level of $\alpha = 0.05$ and doing a χ^2 test we get for the azimuth data a χ^2 value of 29.74 which is above our Threshold value 11.07 with $df = 5$ so according to our test it is significantly different.

Doing a χ^2 test for the zenith data a χ^2 value of 2.24 which is below our Threshold value 7.81 with $df = 3$ so according to our test it's not significantly different.

But since the χ^2 test so sensitive to bin width and which seed we draw our random numbers from since we only have 100 points I don't really trust it be a good test in this case. Therefore i will also peform a KStest which is not dependent on the bin width and which seed we draw our numbers from.

Performing a KStest on the azimuthal data with a uniform CDF between 0 and 2π we get a p-value of 0.0013 which is less than our significance level therefore we can conclude that the data is not from a uniform distribution. On the otherhand perform the same test on the zenith data with a uniform CDF between -1 and 1 we get a p-value of 0.12 which is above our significance level so we cannot reject that the data is from

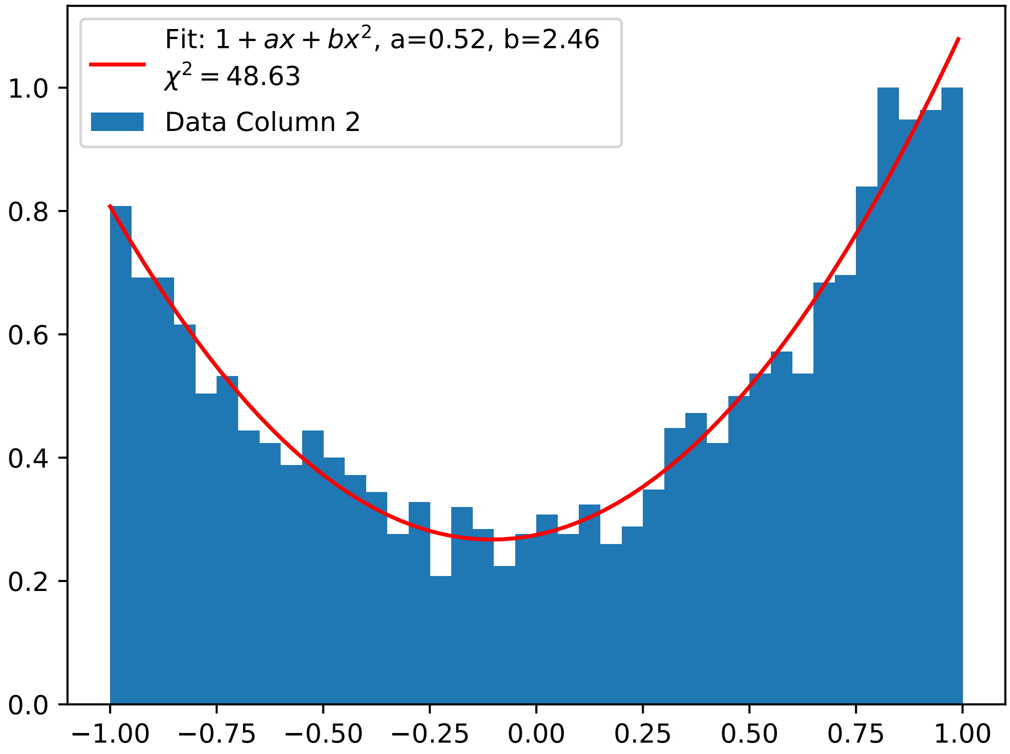

Figure 2: Here we see data from column 2 plotted in a histogram together which its corresponding fit

a uniform distribution.

b)

This time i will generate 20% of the points from a uniform distribution between 0.225π and 0.55π for the azimuth angle and from 0.3π to π for the zenith angle. The remaining 80% will again come from an isotropic distribution. This is H_A.

For H_B i will generate 15% of the points from a uniform distribution between 0 and π for the azimuth angle and from 0.5π to π for the zenith angle. The remaining 85% will come from an isotropic distribution.

Calculating the p-values using a χ^2 test we see that the only hypotheses we cannot reject with a significance level of 0.05 is hypothesis A. The results are summarized in table 1

	$P_{azimuth}$	P_{zenith}
H_A	0.0501	0.131
H_B	0.0297	0.000965
H_{iso}	0.0000166	0.525

Table 1: A Table with the p-value for azimuth and zenith data for the different hypothesis

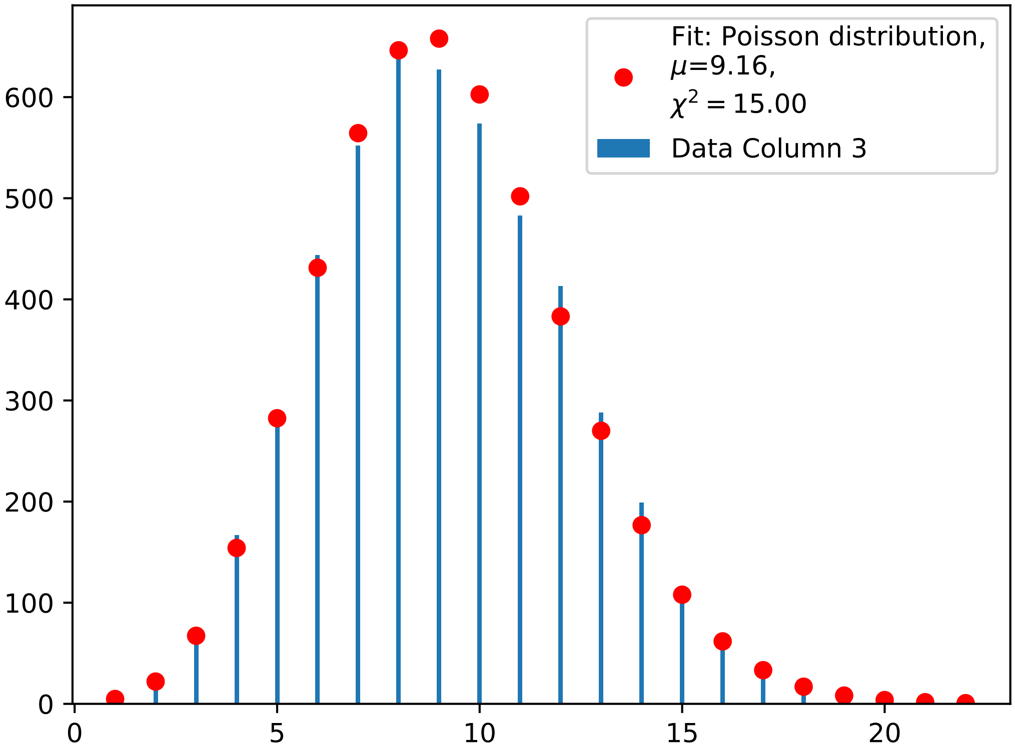

Figure 3: Here we see data from column 3 plotted in a histogram together which its corresponding fit

Problem 3

a)

To solve this problem we sort our data file and find the index of the sorted array where $(1 - 0.0455) * 100 = 95.45\%$ data is below and read the value at that index. In this case it would be index 2864 which corresponds to the value 15.700827 which would be our one-sided p-value threshold.

Calculating the threshold value for a χ^2 for 5 degrees of freedom we get 11.31 which does not match the threshold for the bootstraped data.

b)

Since number attempts for each climber follows a poisson distribution we can use the Compound Poisson distribution theorem to figure out what the most likely number of attempts is. The theorem says you can just add the 4 mean values for each climber. So calculating a poisson distribution for $\mu = 16$ we get the most likely number of attempts is 15 or 16.

The likelihood value for $p = 0.35$ is 2.018. To estimate his posterior probability of succeceding in one given attempt we calculate the mean

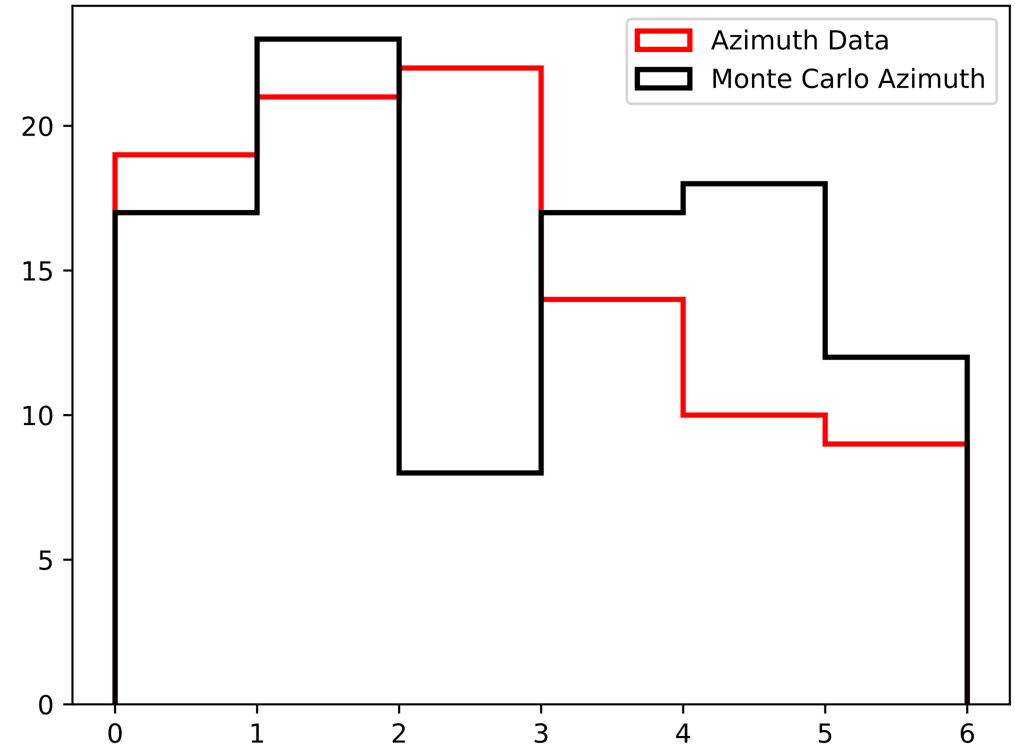

Figure 4: Histogram of 10000 generated points from a uniform distribution and given azimutal data where the histogram is normed

$$\langle x \rangle = \int_0^1 xposterier(x)dx \tag{3}$$

we get $\langle x \rangle = 0.65 \pm 0.04$ The prior, likelihood and posterior is plottet in figure 6

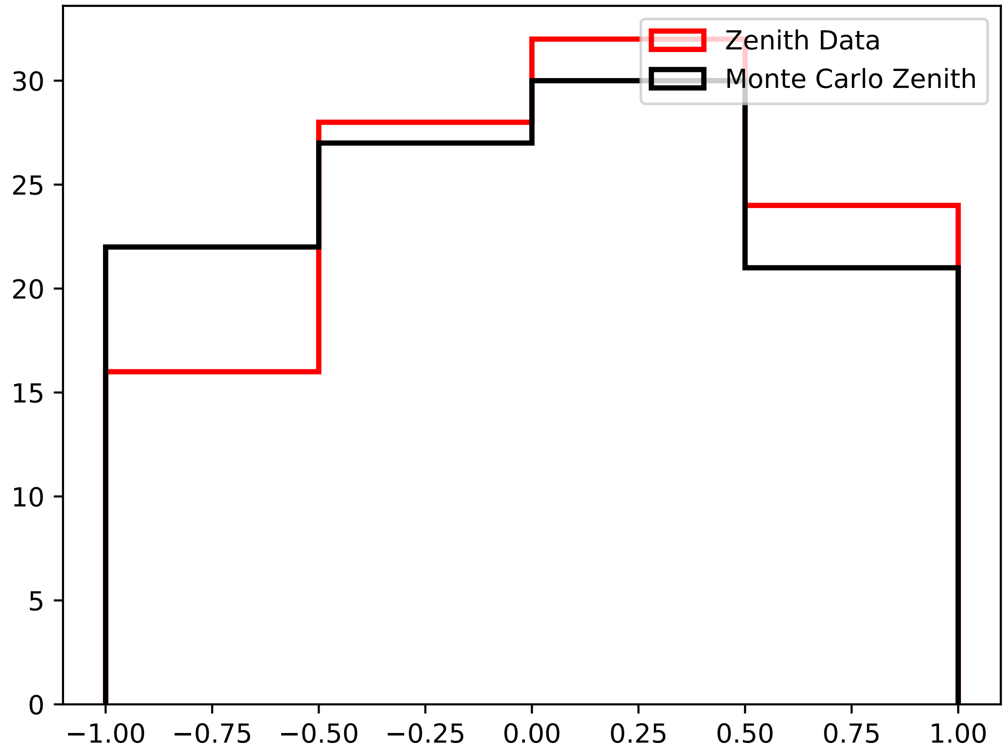

Figure 5: Histogram of 10000 generated points from a uniform distribution and given zenith data where the histogram is normed

Problem 4

a)

XGBoost was used to classify the no-show label. See figure 7

b)

In figure 8 the features are ranked according to the most important to the model where the most important features are placed on top. One way you can test whether or not you begin to overtrain is to look at the error rate of the tested data vs the training data. If your error rate on the tested data increases while your error rate on the training data decreases it's a clear sign that you begin to overtrain and you should consider decreasing the complexity of your model for example in adaptive boosting you would reduce your tree depth or number of trees.

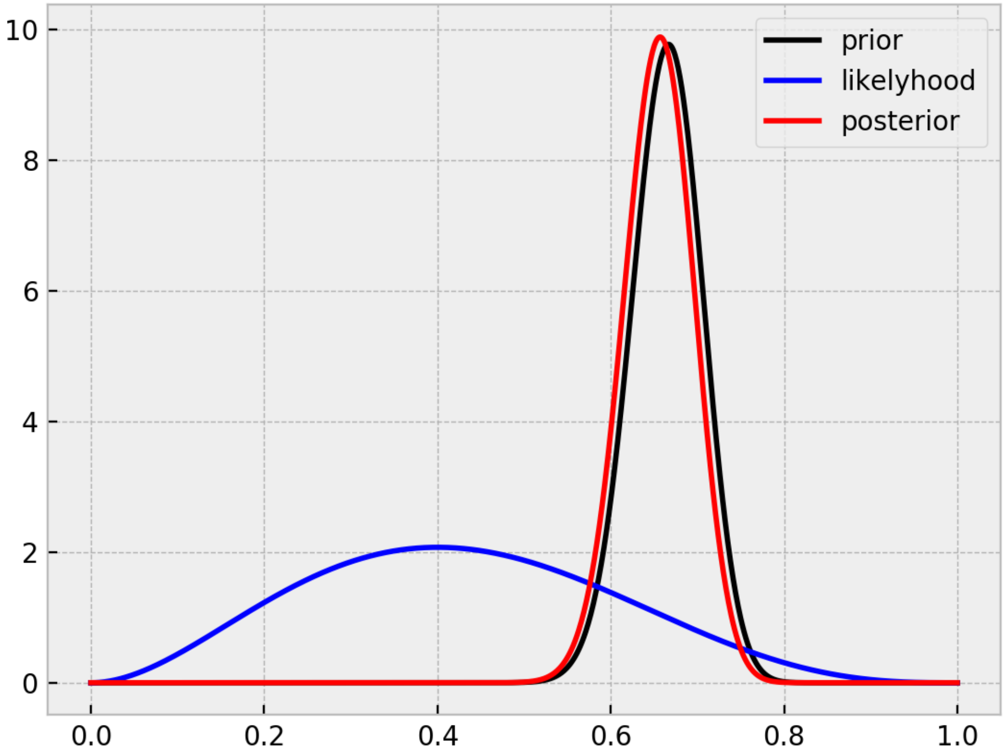

Figure 6: prior, likelihood and posterier plottet

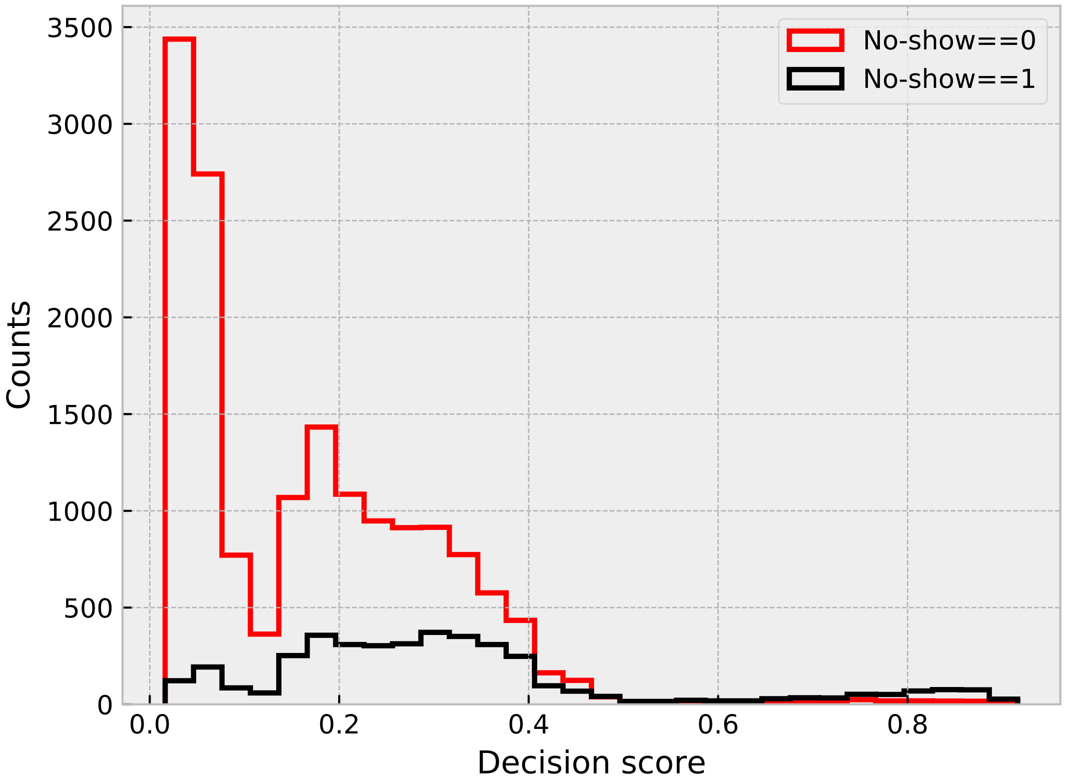

Figure 7: Histogram showing the distribution of the decision scores

1 Problem 5

Doing a 3 dimensional Bayesian nesting we get the best fit parameters to $\theta_1 = 12.35$, $\theta_2 = 6.78$ and $\theta_3 = 0.79$. The result from the bayesian nesting is shown in figure 9.

Doing a raster scan should give us an overview of the likelihood landscape. Therefore we would expect the nesting model to find these maxima. Comparing figure 9 with 11 we see that there should exist some maxima around 0 and 12.5 with indeed the nesting model finds.

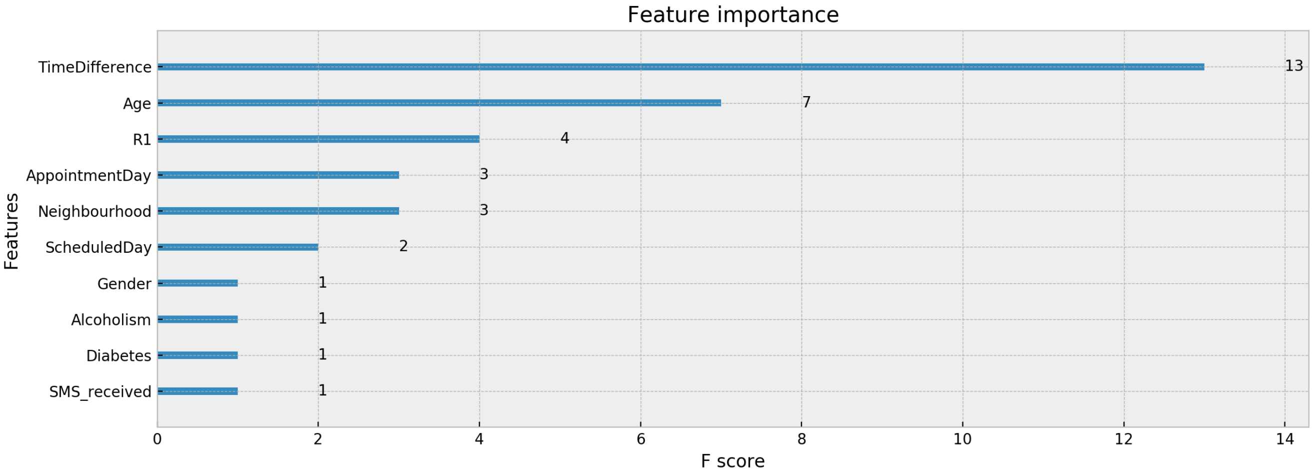

Figure 8: This plot ranks the most important features used by XGBoost

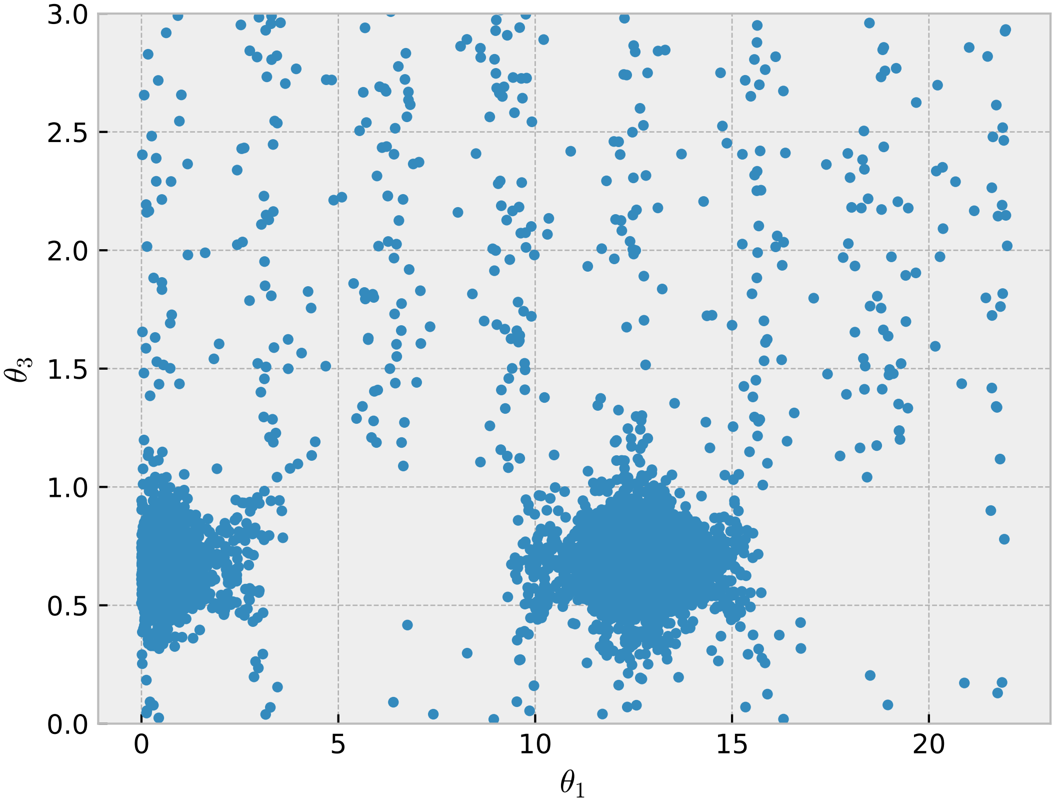

Figure 9: scatter point plot for θ_3 vs θ_1

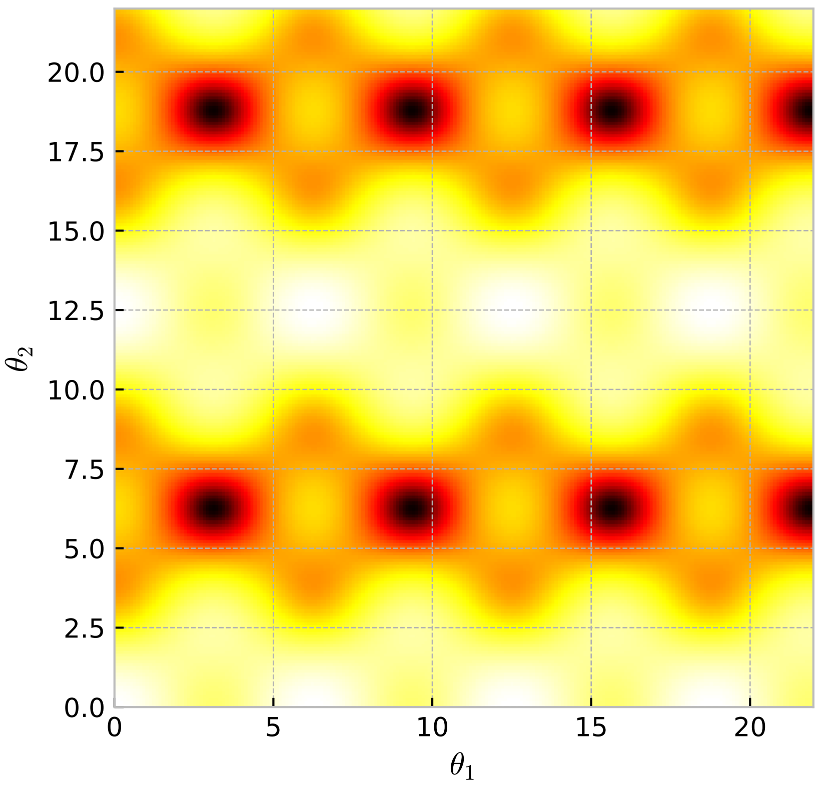

Figure 10: Raster scan for θ_2 vs θ_1 with θ_3 fixed

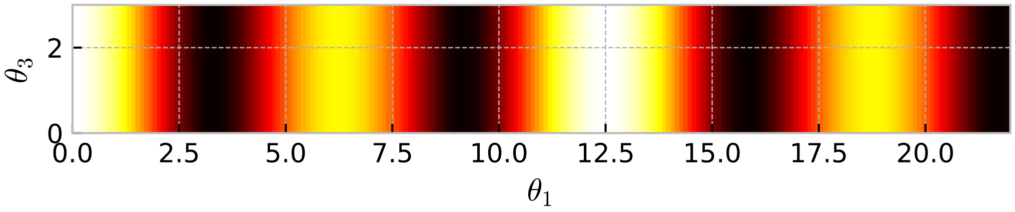

Figure 11: Raster scan for θ_3 vs θ_1 with θ_2 fixed