

Computational Cognitive Neuroscience Assignment, Computational modelling of behavioural data

s1424164

March 20, 2019

1 Part 1

1.1 Task a

Please see Figure 2 and 1.

1.2 Task b

Please see Figure 3. For the sake of space, the labels of the axes are omitted. The histograms plot count (on the y axis) against time (on the x axis). The paths plot time (in seconds, on the x axis) against Wt (on the y axis). When the z parameter is reduced, with high noise, the accuracy reduces to 74% and the reaction times increase. This is because a reduced z parameter pushes it closer to the H- decision boundary. However, with no noticeable noise, there is no reduction in accuracy, but still an increase in reaction times. For the modification in the a parameter, the accuracy remains the same in both cases (99%) (with noise and less noise), though the reaction times definitely reduce. For the v parameter, there is both an increase in accuracy and decrease in reaction times when sufficient noise is added, for the lower v, but only an increase in reaction times when the noise is lower. For an increased S parameter, there is a reduction in reaction times but also a reduction in accuracy to 50%, which essentially makes this a random process.

1.3 Task c

In order to explain the first finding that depressed participants showed an increase response time, but no difference in accuracy, it is sufficient to hark back to part 1b where a modification in the a parameter, with an increased s parameter, has precisely this effect and so no further demonstration is needed. Depressed participants would therefore have an increased s parameter under this model. Depressed participants would also have a higher variance in the reaction times under this interpretation too. This can be seen on the second row of Figure 3.

Without knowing the specifics of the experiment performed, it is difficult to infer which parameters might have changed. However, in order to explain the second finding that depressed participants showed similar response times, but a reduced accuracy, it would be sufficient to refer to the modification in the z parameter, with an increased s parameter, which appears on the first row of Figure 3. Here, a value of z closer to the decision boundary of H- (which almost looks like a negative bias on results), reduces accuracy and creates a bimodal distribution for blue parameters in terms of reaction times, though the mean of the reaction times distribution remains the same.

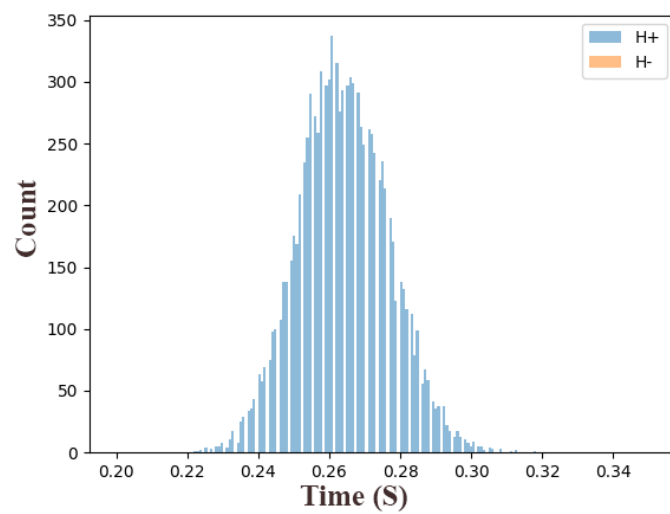


Figure 1: Reaction time distributions.

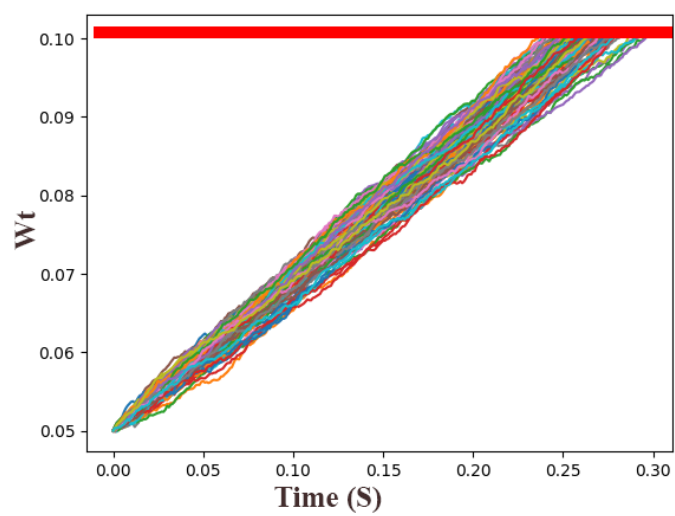
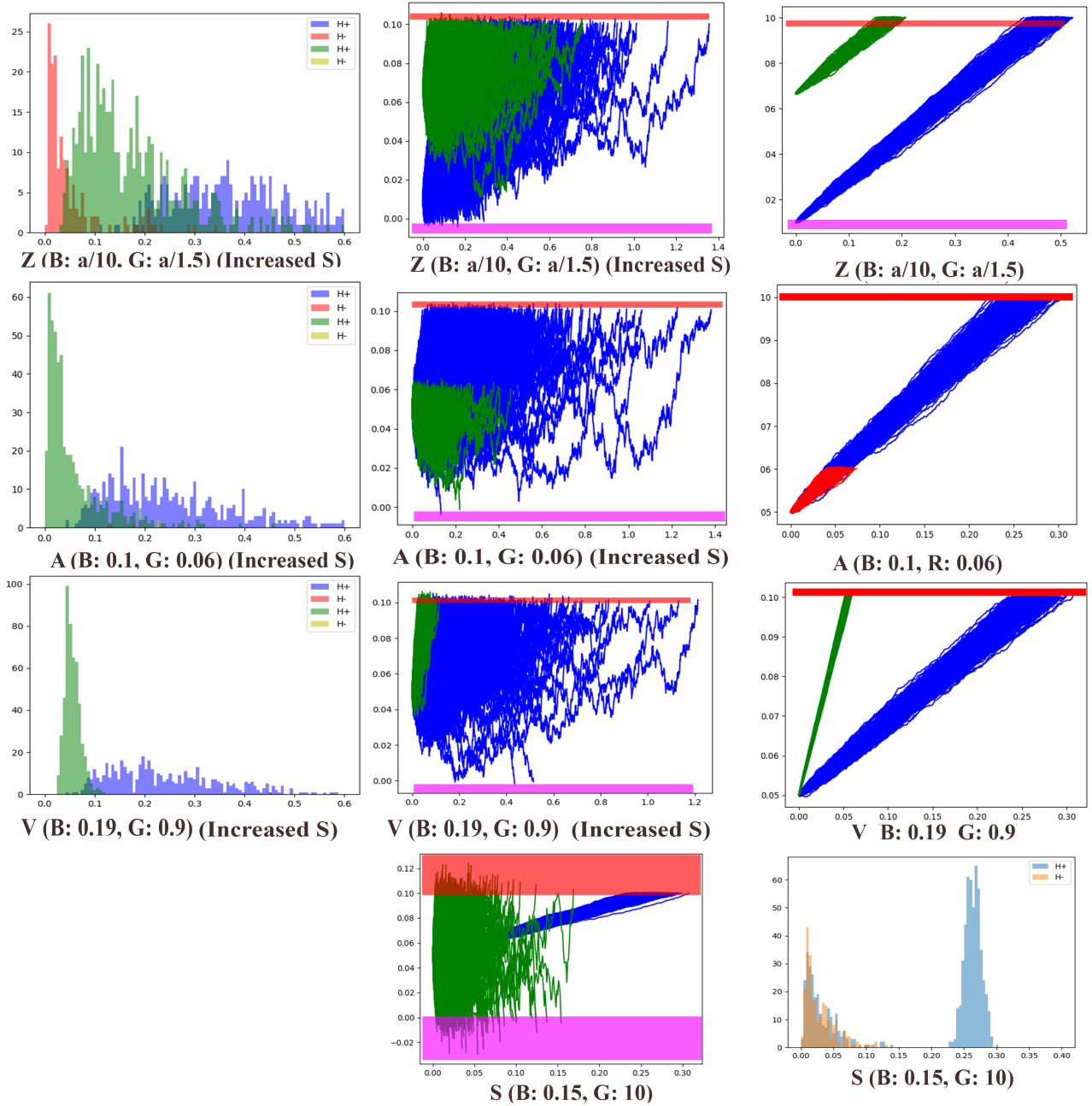


Figure 2: One thousand simulated paths.

Figure 3: This figure shows the differences in accuracy and reaction times for different parameter settings, that deviate from those specified in part a. At each trial 1000 simulations took place. The correct hypothesis is H_+ , which is denoted by paths that reach the top line in red. Depending on the noise parameter s , which is added into the process, different parameter settings have different effects. This is because a substantial increase in noise overrides the effect of v . The first three rows depict on the far left, the histogram of responses. In green, are the correct hypotheses coming from the green parameter settings from the paths on the immediate right (those with lower noise (s)). In blue are the correct responses also from the blue parameter settings. In red, are the incorrect responses coming from the blue parameter settings from the paths and in yellow the incorrect responses coming from the green parameter settings. The red and pink lines display the boundaries for H_- and H_+ respectively. On the second row, due to the overlapping paths, for clarity, the reduced A is shown in red. On the bottom row, the histogram has colours orange and blue for both the green and blue parameter settings. However, it is clear that the green parameter settings induce the inaccurate blue and orange cluster at the left of the rightmost histogram.



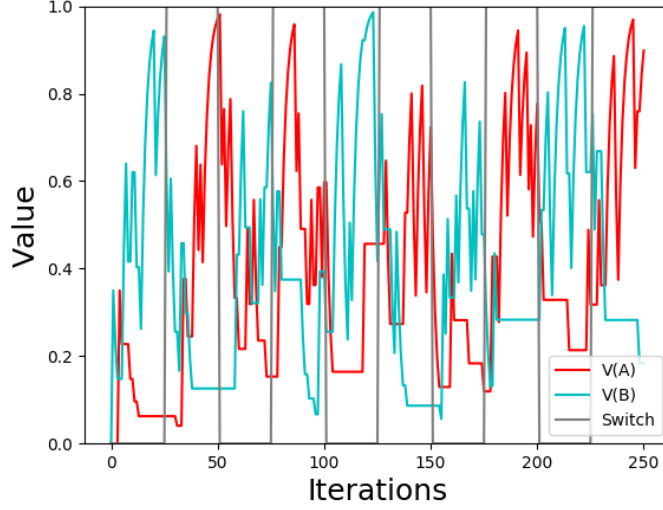


Figure 4: This shows that during the odd indices of the intervals (starting from 1), the value of B increases and in the even interval indices the value of A increases. It is interesting to note that the red line shows a steeper increase during changes, which is to be expected as the difference in reward probabilities is greater when the probability of reward for A is greater than B, than when B is greater than A.

2 Part 2

2.1 Task a

To see the evolution of values please see Figures 4 and 5.

The expected number of received rewards can be calculated by taking the probability of the reward to be 0.5 and then using the definition of expectation over the random variable of rewards. This works out to be $(125 * (0.5 * (0.45 * 1) + 0.5 * (0.7 * 1) + 0.5 * (0.7 * 1) + 0.5 * (0.3 * 1))) = 134.375$. The value 125 is used here to account for the iterations being divided into slots of 25.

2.2 Task b

Please see Figure 6, with the caption for an explanation. Figure 7 details the effect of different temperature values on the calibrating effect of two Gaussians, using temperature scaling, a method called Boltzmann scaling, first introduced to mainstream machine learning in [1].

2.3 Task c

The NLL for the second participant is 127.2.

2.4 Task d

The different initialisations I used were $[0, 0]$, $[0.5, 5]$, $[1, 10]$. The differences for all parameters, for all settings were in the order of $1e-08$, which is smaller than the tolerance of the minimisation convergence criterion, which infers that the minimisation process is potentially convex, with a single global optima. However, unless mathematical analysis is performed, this cannot be conclusively established.

Figures 8 and 9 show that depressed patients have a higher, but more variable, learning rate, but also tend to favour other options more easily than the control group, who learn slower, but once they have learned, they stick with the learned choice more. It is unclear why this is the case. For this analysis, the

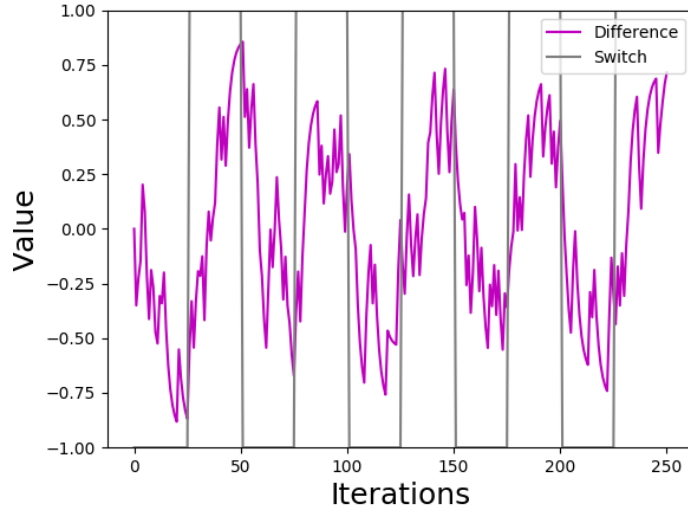


Figure 5: This shows that the difference oscillates in accordance with the differences in reward probabilities. The peaks and troughs generally are found at the value of 0.75. The decreases occur when B has a greater reward probability and vice versa for when A has a higher reward probability.

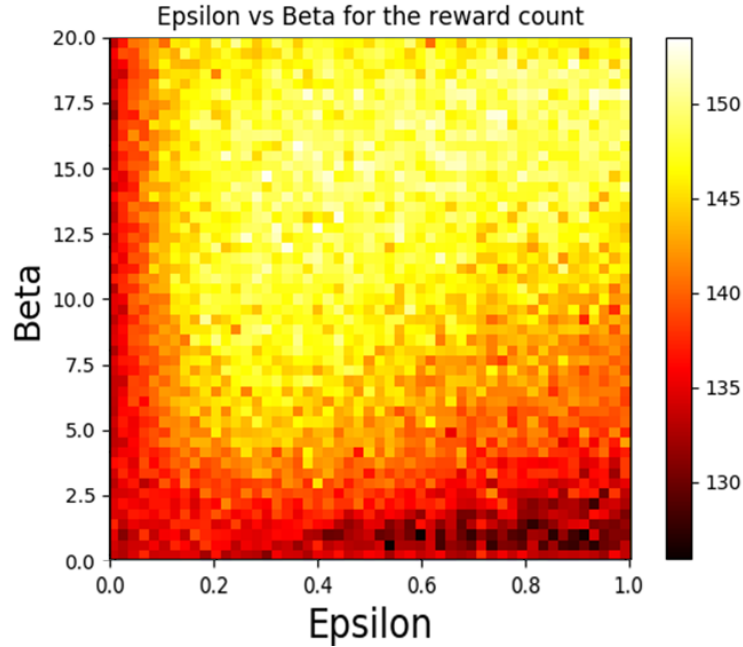


Figure 6: The beta parameter scales the softmax function in a framework known as temperature scaling, with a lower value having the effect of reducing the gradient of the Gaussian. More explicitly, the beta parameter recalibrates the decision boundary between the two Gaussian clusters representing choice A and choice B. A lower value of the beta parameter means that at each iteration, a greater proportion of the probability mass is distributed to the choice with the lower value than at higher values of beta. The epsilon parameter on the other hand determines how fast a given reward re-weights the value of a choice. A low value of beta, means that regardless of the values of the choices, the probability is around 0.5 for both choices. Clearly, this reduces performance to average performance. This also occurs for epsilon values of zero where no learning occurs. High values of epsilon and beta values yield better and better results, as the learning improves and is more responsive to changes in values. It is interesting to note that values of epsilon around 0.9 change rewards in response to increasing beta more slowly than epsilon values of 0.4.

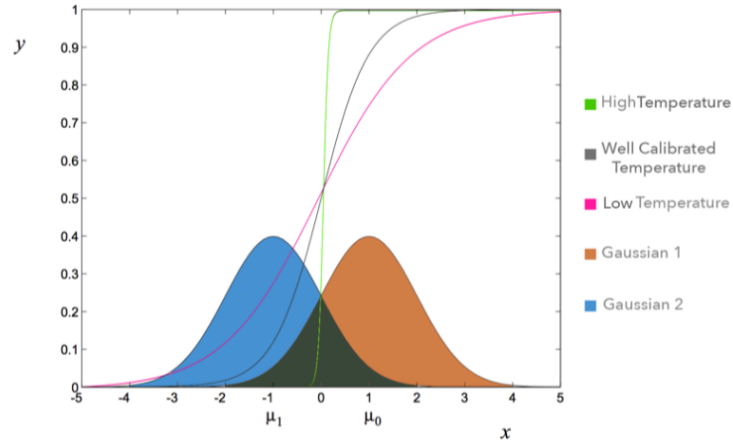


Figure 7: This figure demonstrates the effect of different inverse temperature values on Gaussian calibration.

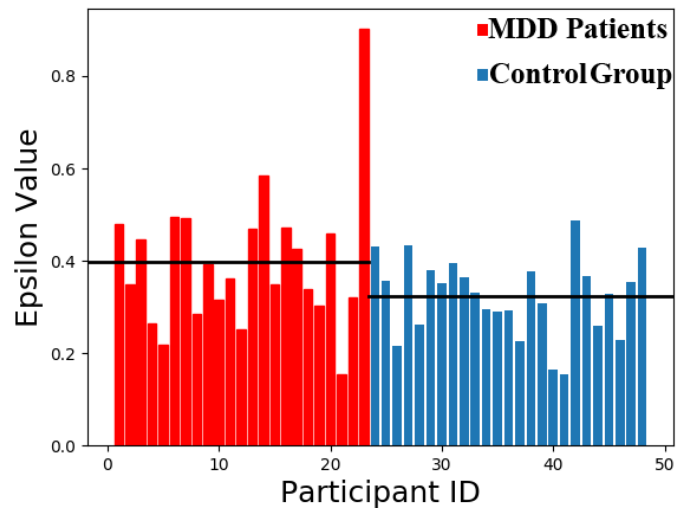


Figure 8: Here, it can be seen that the epsilon values of the MDD patients are higher than the control group. The epsilon values of the patients with MDD also have a higher variance. The black line represents the average value of each group.

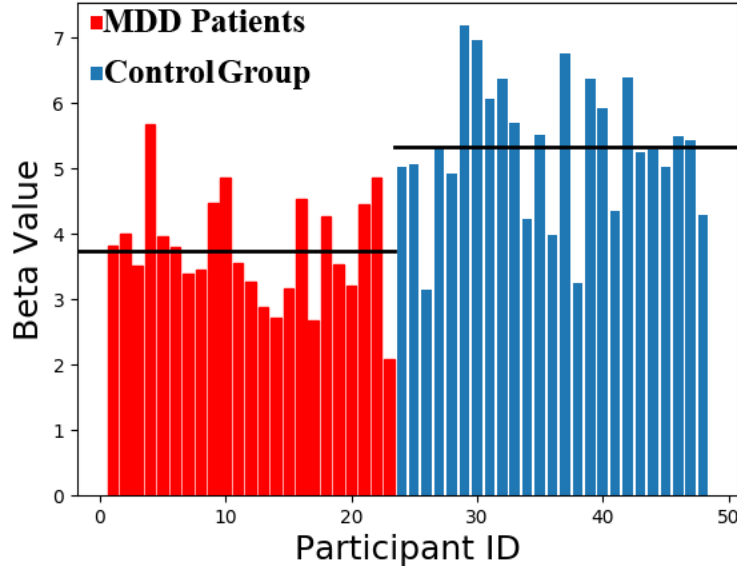


Figure 9: Here, it can be seen that the beta values of the control group are higher than the MDD patients. The black line represents the average value of each group.

Spearman's Rank correlation will be used for analysis, but for completeness Pearson's correlation coefficient will be included. This is because, normal tests revealed that the data is not normally distributed.

Correlation	Pearson	Pearson P Value	Spearman Rank	Spearman P Value
Across Group	-0.36	0.01	-0.24	0.09
Between Group MDD	-0.60	0.003	-0.46	0.03
Between Group Control	0.18	0.38	0.16	0.45

It can be concluded that there is a slight negative correlation for both groups between epsilon and beta. However, this masks the individual differences. There seems to be a moderately strong, significant, negative correlation between beta and epsilon for the patients with MDD. For the control group, there is an insignificant, though positive correlation between epsilon and beta. Across groups, there is a weak negative correlation between beta and epsilon, which is significant for the Pearson correlation but not for the Spearman Rank correlation.

2.5 Task f

Parameter	t-statistic	p-value	d.f
ϵ	2.05	0.048	34.1
β	-5.79	0.000000656	44.78

Welch's student t-test was performed which measures whether there is a difference in expectation, without assuming a difference in variances. Here, H_0 is the hypothesis is that both ϵ and β are equal to each other across groups. The alternative hypothesis H_1 is that these two parameters are not equal to each other, i.e. that depression could be related to changes in reward learning (i.e. an altered learning rate) or some change in decision making (inverse temperature). Both p-values are below the 5% critical value, which leads me to reject H_0 , for both parameters. The t-statistic for them both show that they are anywhere from two to five standard deviations away from H_0 . The degrees of freedom here, represent the amount of data that is independent once we know the number of data points corresponding to the degrees of freedom. Therefore once we know 34 values in the case of ϵ we know all 48. The degrees of freedom in the case of ϵ are lower than $n - 2$ which tells us that potentially the tails of our distribution are fatter than

usual. β on the other hand has more mass in the centre of the distribution. If the data were real, then we would reject the null hypothesis and hypothesize that epsilon and beta both have significantly different averages for MDD patients and the control group.

2.6 Task g

The mean and variance of parameter values are described below. The mean and variance I chose to use for this part are the **Total** values. The other values here are listed for later analysis.

Parameter Values	Mean (3 sf)	Variance (3 sf)
ϵ Total	0.359	0.016
ϵ MDD	0.397	0.022
ϵ Control	0.323	0.00697
β Total	4.57	1.51
β MDD	3.75	0.653
β Control	5.33	1.09

I chose to sample fifty sets of epsilon, beta pairs, a value above the number of degrees of freedom of the data and high enough such that accurate correlation values were achieved. Figures 10 and 11 demonstrate the sampled values. Here, the Pearsons correlation coefficient can be used as the data is normally distributed. Spearmans Rank correlation coefficient is included for completeness. Figures 12 and 13 show the sampled values against the recovered values. The correlation values were as follows:

Correlation	Pearson	Pearson P Value	Spearman Rank	Spearman P Value
Epsilon	0.259	0.0690	0.220	0.0124
Beta	0.897	1.22e-18	0.896	1.66e-18

This parameter recovery does not meet my expectations. I would expect the parameters to be highly correlated. As I cannot be certain of the normality of the underlying variables, I will use Spearman's Rank correlation in the following analysis. It is clear from figures 14 and 15, that as the number of trials performed increases, the correlation goes up for both beta and epsilon. However, if only ten sets of parameters are used, then the increase decreases for epsilon and almost flatlines for beta. This shows parameter recovery benefits from many sets and many trials. However, the number of sets seems to have less of an effect than the number of trials.

2.7 Task h

The two models have been fitted and the results are shown in Figures 16 to 20. For model 2, the results are exactly identical to the results of model 1. This is because the parameter settings are equivalent to each other, a notion that can be intuited, although not formally proved via the following inductive reasoning.

If we equate the exponential terms, this relationship can be seen.

For $V_i^t = 0$ as a basis step.

$$\exp(\beta * 0) = \exp(0)$$

Then for the next step this becomes

$$\begin{aligned} \exp(\beta * (0 + \epsilon * (r^t - 0))) &= \exp(0 + \epsilon * (\rho * r^t - 0)) \\ \exp(\beta * \epsilon * r^t) &= \exp(\epsilon * \rho * r^t) \end{aligned}$$

which are identical. As V_i^{t+1} is defined in such a way we are done. Therefore, whichever values of beta are found will also be those found by rho.

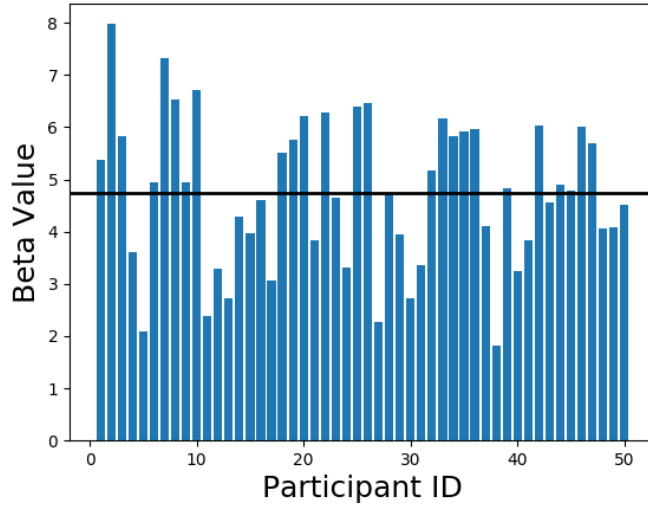


Figure 10: Task G) Here, beta values are plotted sampled from a normal distribution with the Total parameters.

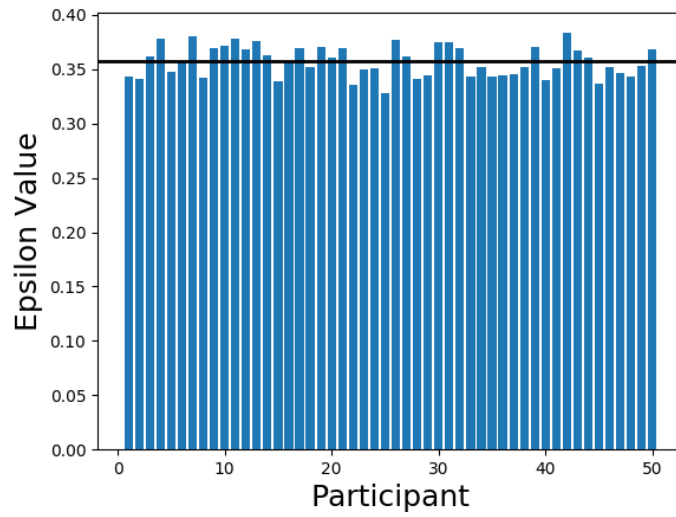


Figure 11: Task G) Here, epsilon values are plotted sampled from a normal distribution with the Total parameters.

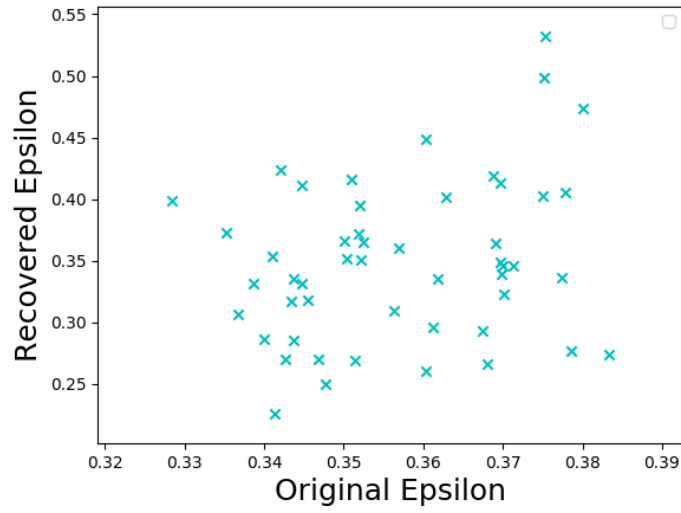


Figure 12: Task G) Here, the recovered epsilon parameters are plotted against the generated (or original) epsilon parameters.

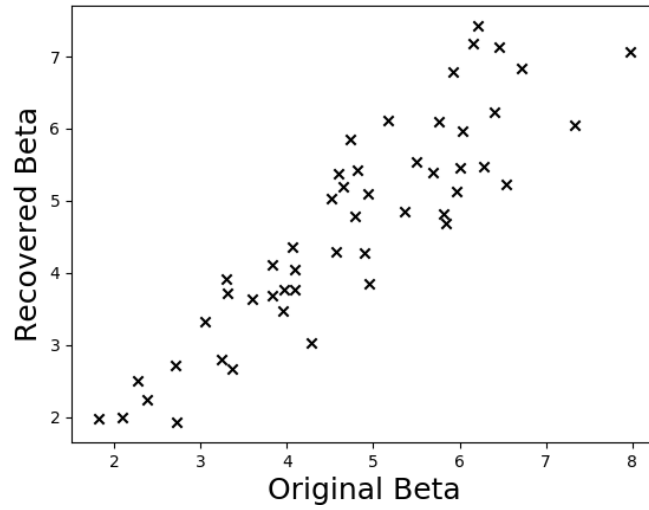


Figure 13: Task G) Here, the recovered epsilon parameters are plotted against the generated (or original) epsilon parameters.

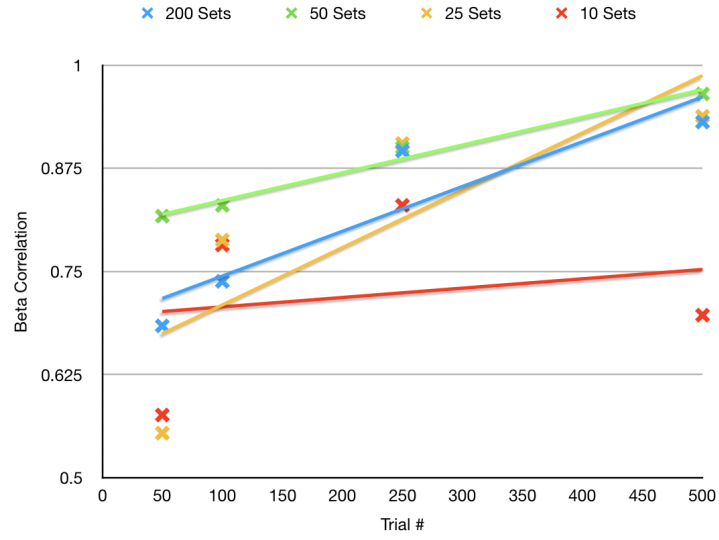


Figure 14: Task G) Here, different trial numbers and set sizes are shown for the beta parameter.

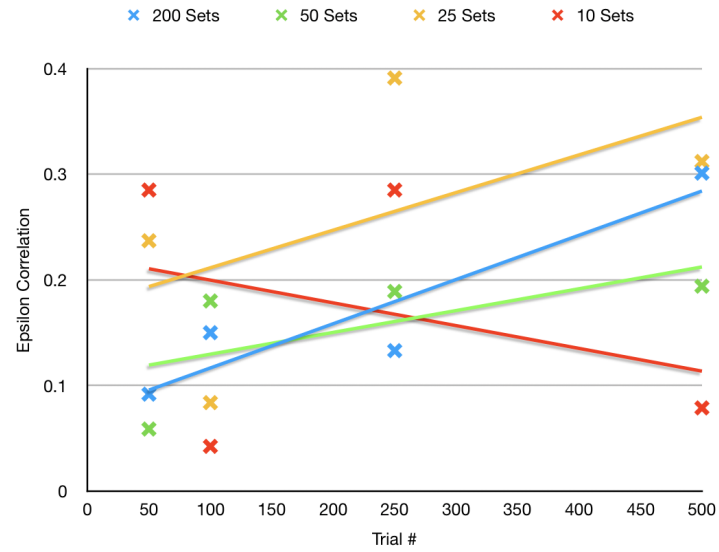


Figure 15: Task G) Here, different trial numbers and set sizes are shown for the epsilon parameter.

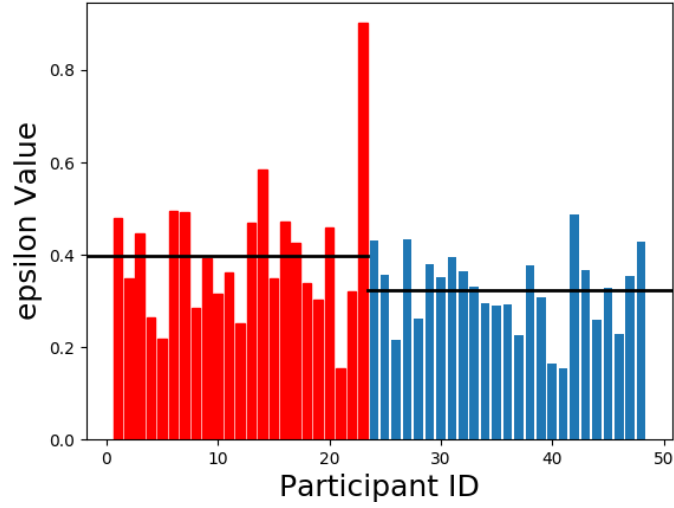


Figure 16: Task H) Epsilon Model 2

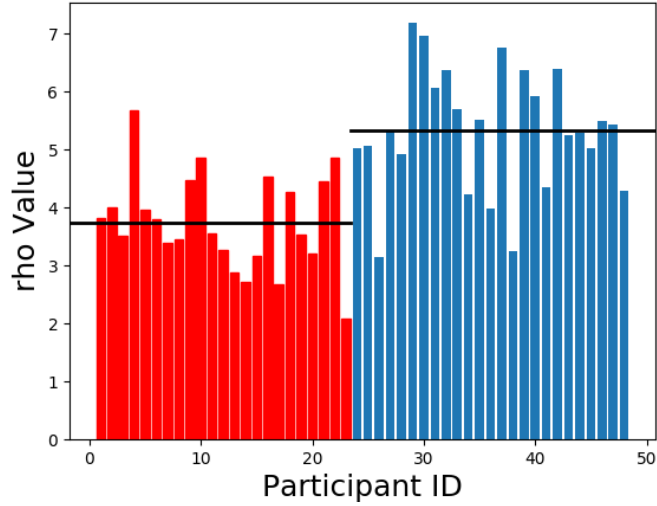


Figure 17: Task H) Beta Model 2

For model 3, these are also identical to model 2 and to model 1. The results are as expected, because even though the initial parameters are set to the same value, as along each dimension, the error surface which dictates the element that minimises the negative log likelihood is convex as established earlier. Therefore the elements of v_0 do not affect beta and epsilon. I cannot infer why these parameters are set as they are, which renders them surprising to some degree.

2.8 Task i)

The AIC and BIC values are listed as follows:

Model	AIC	BIC
1	10743.8449913	11081.9052394
2	10743.8449913	11081.9052394
3	10754.4645429	11261.554915

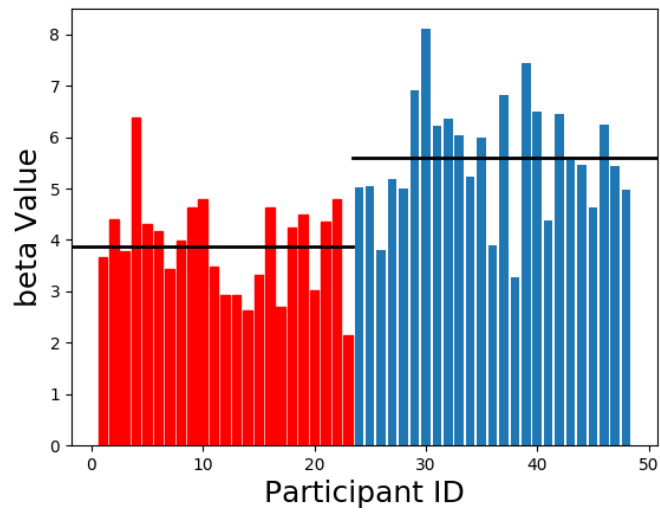


Figure 18: Task H) Beta Model 3

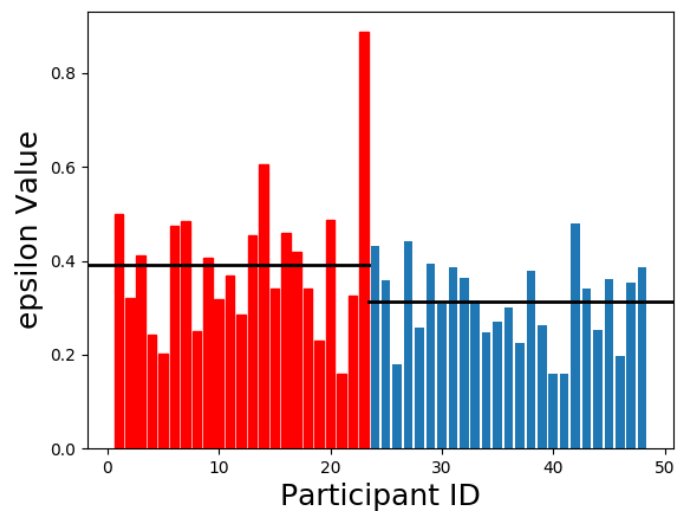


Figure 19: Task H) Epsilon Model 3

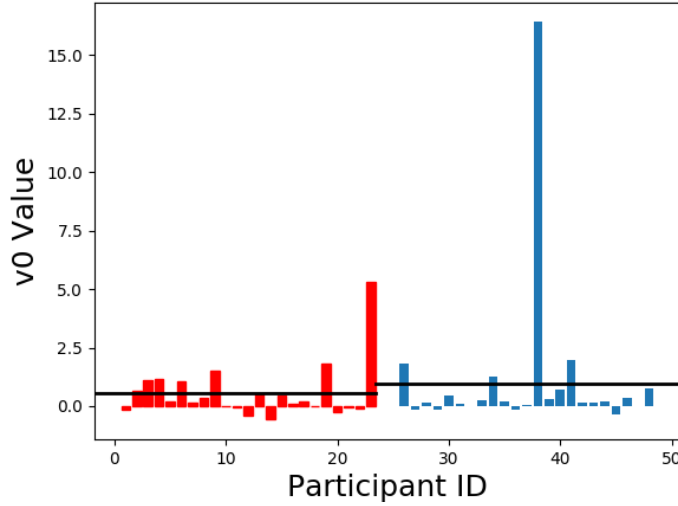


Figure 20: Task H) V0 Model 3

The model I would choose would be either model two or model one as they both minimise the AIC, the BIC and the AICc. The third model is less likely to minimise information loss over the data. The third model also features an extra parameter, which is likely unnecessary as it has the effect of increasing the AIC. The AICc is an adjusted measure used to test for overfitting, which features a penalty for extra parameters, and is also minimised over models one and two. The BIC is also minimised which penalises extra parameters more than the AIC, yet it still prefers model two and model one.

2.9 Task j)

Figures 21 and 22 show the confusion matrices. The approach that I took was to, for each model, first to fit the model to the data, then to generate parameter sets from the fitted data. This generated data was sampled from a normal distribution, with the same mean and variance as the fitted parameters. I then re-simulated data from that model with these new generated parameters. From this data, I fitted all three models and calculated the AIC scores for each model. I did not choose the scores to be linear combinations of AIC and BIC, due to the potential complexity involved in ascertaining the relative weights. If the data was generated from model one, but model three has the lowest AIC score for the data, then the data point was said to be a ground truth instance of model one, but predicted to be model three. This creates a framework for constructing the confusion matrices. Comments on these matrices are featured in the captions of the matrices.

2.10 Task k)

This analysis was performed in a sensible way, i.e. we started with a hypothesized model, then we did group analysis, before recovering parameters and trying additional models and comparing them. In task f), it was concluded that different groups were found in the data. The models that we found to be the best were the same ones that the analysis in part f) was concluded on, therefore my interpretation of the results have not been changed. My final conclusions are that depression is related to a change in learning rate and a change in inverse temperature with respect to the models.

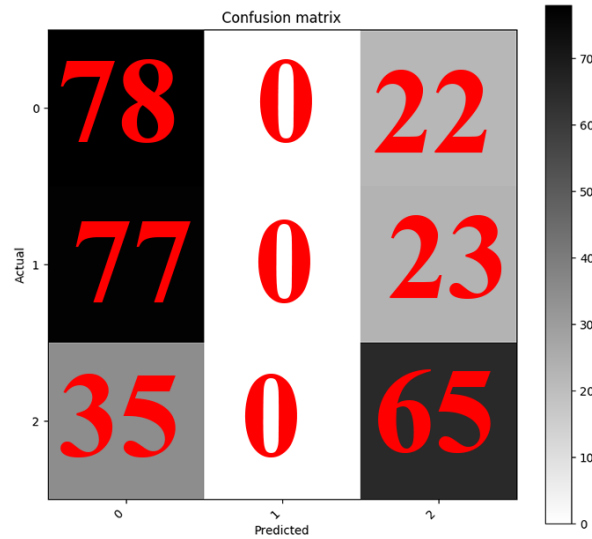


Figure 21: Task G) This details the confusion matrices for values that have been rounded to two decimal places. On the y-axis are the actual values, with the top row (0) being model 1, the second row as model 2 and the third row as model 3. On the x-axis are predicted values, with the leftmost column being model 1, the middle column being model 2 and the rightmost column being model 3. It is interesting to see that because model one and model two have the same performance, as the min operator chooses the first element in the array when doing comparisons against the models, and as model one always comes before model two, the model never predicts model two as the best. This means that model twos results are in some sense 'captured', by model one. Model three sometimes minimises the AIC on model ones data, though model one minimises model threes data more often.

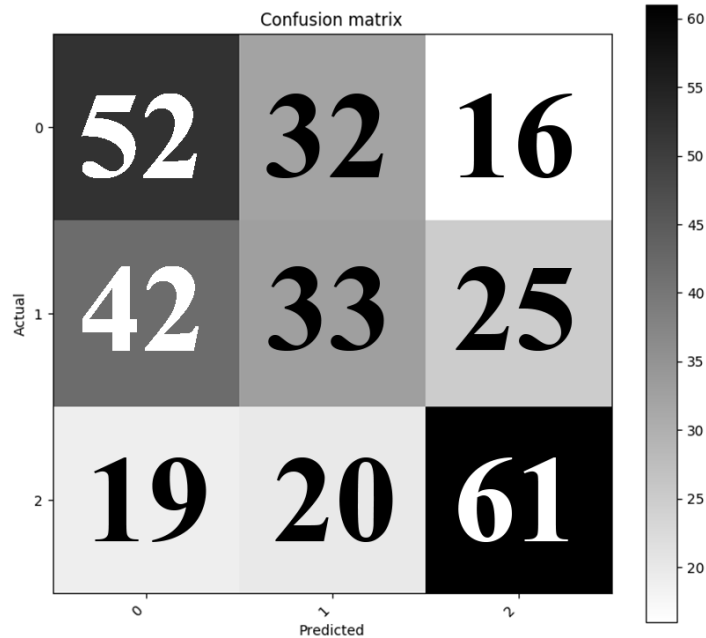


Figure 22: Task G) If no rounding is used, because of small differences in decimal places, model two is often chosen as the minimum over model one.

References

- [1] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.