Probability estimation simulations

Chanwoo Lee, September 27, 2020

1 IQ brain data CV result with the modified algorithm

I perform cross validation on IQ brain dataset with the modified algorithm. Figure 1 shows the cross validation results according to 0-1 loss and the hinge loss. The result is pretty much the same with the result with the previous algorithm. Based on the results, I set rank 4 and estimate conditional probability. Figure 1 shows the results. We can check that the probability is divided into two distinctive groups around FSIQ 120, which is threshold of the labeling. The additional sparse structure on the coefficient might be needed to have better probability estimation.

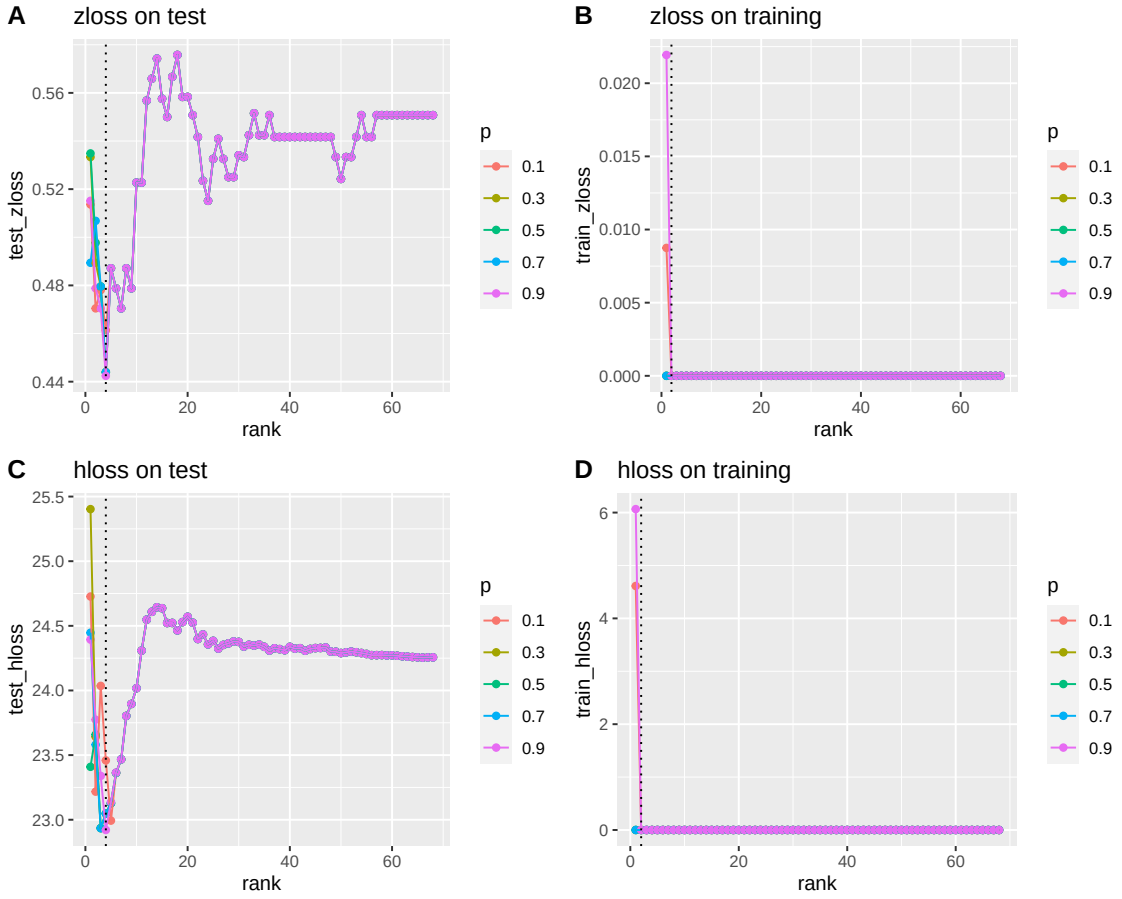


Figure 1: Cross validation results. The dotted vertical lines in A,C are when rank is 4 while in B,D are rank 2.

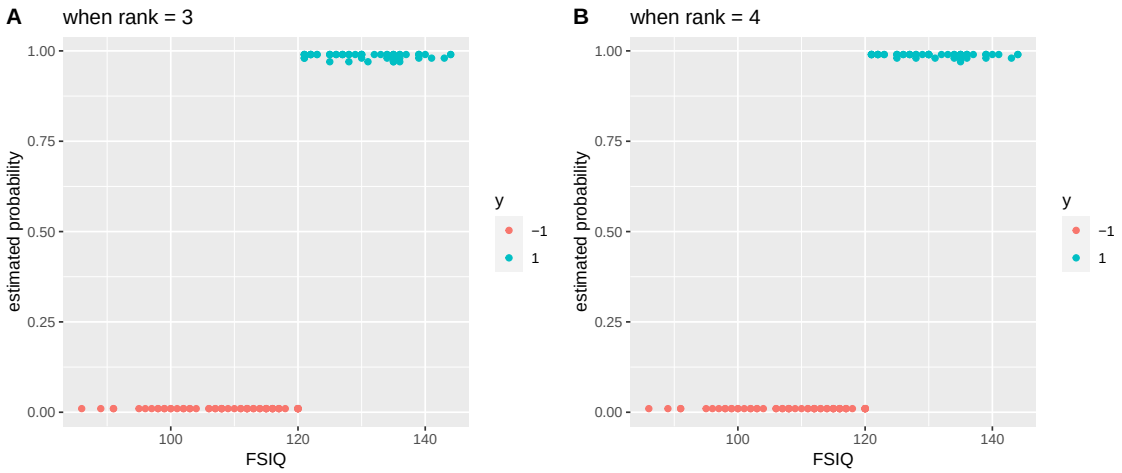


Figure 2: Probability estimation results. the x-axis is FISQ value and y-axis estimated probability.

2 Sanity check simulation

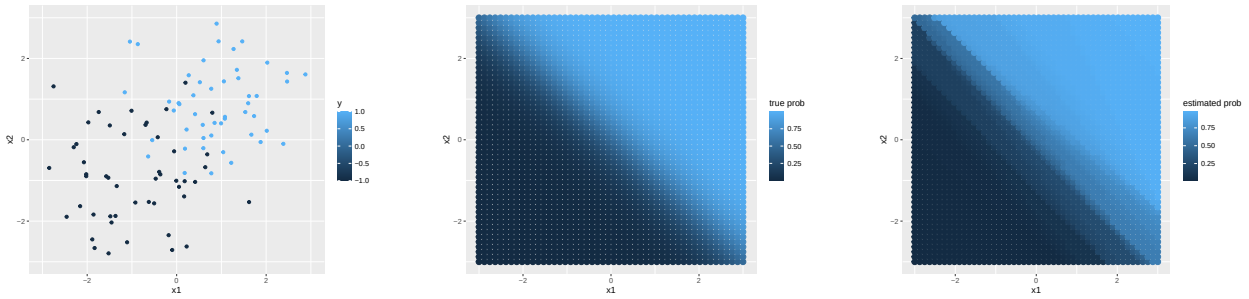
2.1 Probability estimation on a dataset with normal distribution.

I check the algorithm performance based on normal distribution. I set the training data points $\{(x_1, y_1), \dots (x_{100}, y_{100})\}$ based on the following rule.

$$x_i \overset{\text{i.i.d.}}{\sim} \begin{cases} N((1,1)^T, I) & \text{when } y_i = 1, \\ N((-1,-1)^T, I) & \text{when } y_i = -1 \end{cases}$$

Same set-up as in previous note?

Here $y_i = 1$ for $i = 1, \dots 50$ and $y_i = -1$ for $i = 51, \dots, 100$. Figure 3 shows the training data points, true probability based on the rule, and the estimated probability. The result shows that our algorithm successfully estimates true probability.



(a) Training data points (b) True probability (c) Estimated probability

Figure 3: The probability estimation on data points from normal distribution. (a): Training data points, (b): The true probability, and (c): The estimated probability.

2.2 Probability estimation from Bernoulli distribution with logistic link

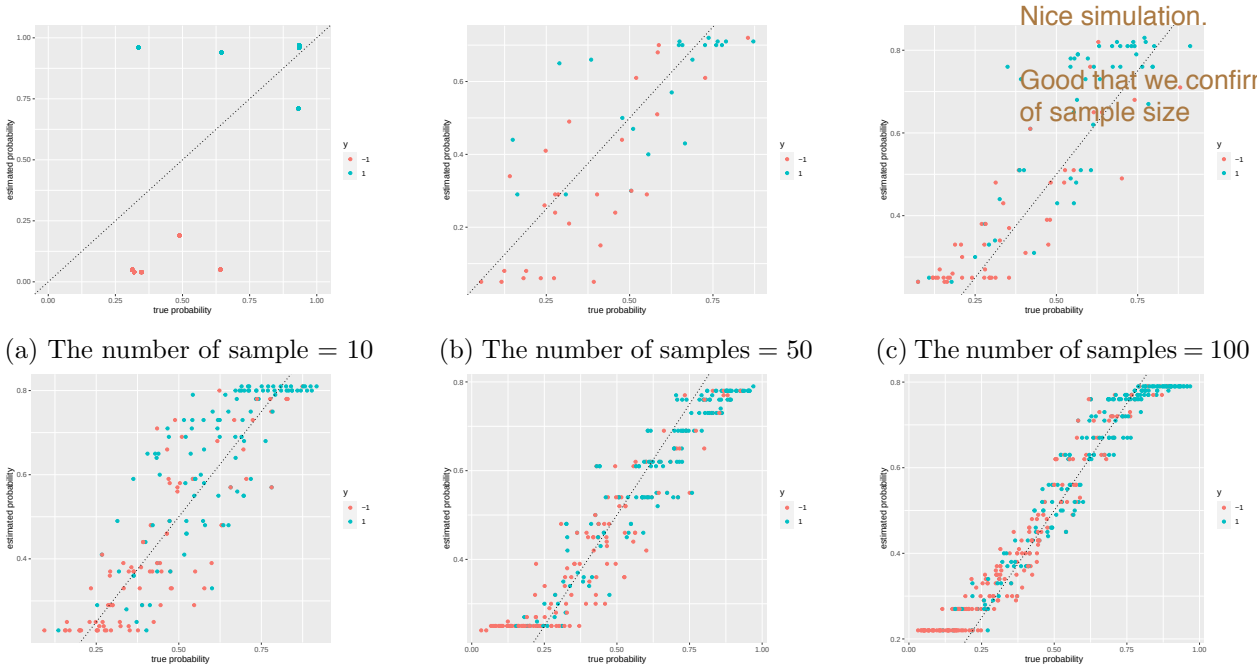
I generated the training datasets whose feature matrices $\boldsymbol{X}_i \in \mathbb{R}^{5\times 5}$ are symmetric binary matrices. I assigned the label responses y_i as

not i.i.d., only independent —> each y_i has different success probability

$$y_i \overset{\text{i.i.d}}{\sim} \text{Ber}\left(\text{logistic}(\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle)\right),$$

how strong is the signal? plot hist(logistic(<B, X_i>))

where $\boldsymbol{B} \in \mathbb{R}^{5\times 5}$ is a rank 3 coefficient matrix. I changed the sample size $N \in \{10, 50, 100, 200, 300, 400\}$. Figure 4 shows that our algorithm estimate the probability more accurately as the number of training data points increases. Notice that the proportion between the number of sample size and free parameter in the brain IQ dataset is similar in the case when the number of sample size $= 10$ in this simulation. Therefore, This simulation suggests that we need more samples or less free parameter by adding sparsity structure on the coefficient matrix to obtain better probability estimation.



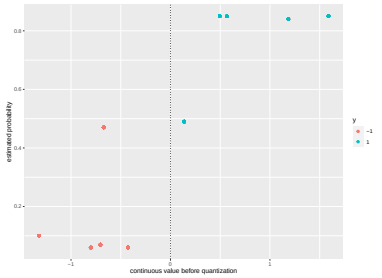
(a) The number of sample $= 10$ (b) The number of samples $= 50$ (c) The number of samples $= 100$

(d) The number of sample $= 200$ (e) The number of samples $= 300$ (f) The number of samples $= 400$

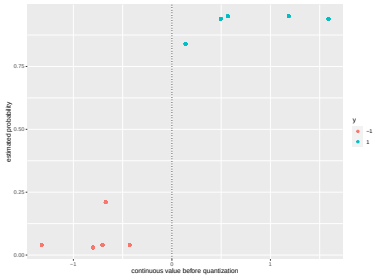
Figure 4: Probability estimation on logistic model. The x-axis is the true probability and y-axis the estimated one. Each small figure shows the output according to the sample size.

Based on the above observaion, I check weather low rank structure helps to improve performance of probability estimation. To make similar proportion in the brain IQ dataset, I set the number of observation as 10 and choose the rank in $\{1, 3, 5\}$ where 3 is the true rank. The first three figures in Figure 5 are based on the case when the label is chosen from quantization at 0. The last three figure are based on logistic model. When rank is 1, it seems to give us better probability estimation results.

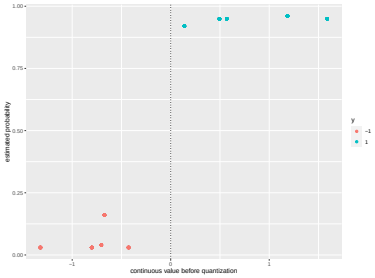
Based on the above result, I tried to use rank 1 or 2 on the brain IQ dataset and check whether the probability estimation results can give us information about true IQ of the subjects. Figure 6 shows that when the rank is 1, the magnitude of the probability gives us better information about true probability than other ranks.



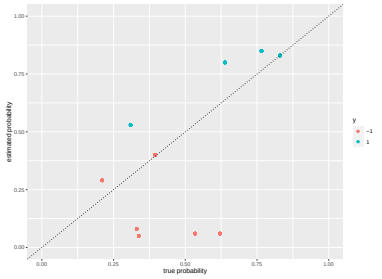
(a) Case 1: Quantized labeling. When the rank is 1



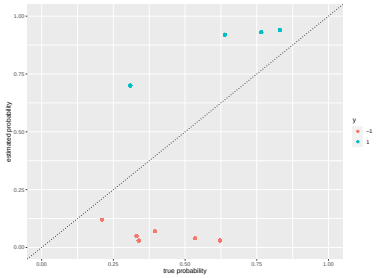
(b) Case 1: Quantized labeling. When the rank is 3



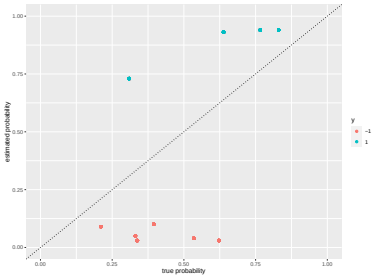
(c) Case 1: Quantized labeling. When the rank is 5



(d) Case 2: labeling from logistic model. When the rank is 1



(e) Case 2: labeling from logistic model. When the rank is 3



(f) Case 2: labeling from logistic model. When the rank is 5

Figure 5: Probability estimation based on quantization or logistic model according to different rank.

could you try n = 100?
Hard to tell the difference between (a)-(c) when n = 10.

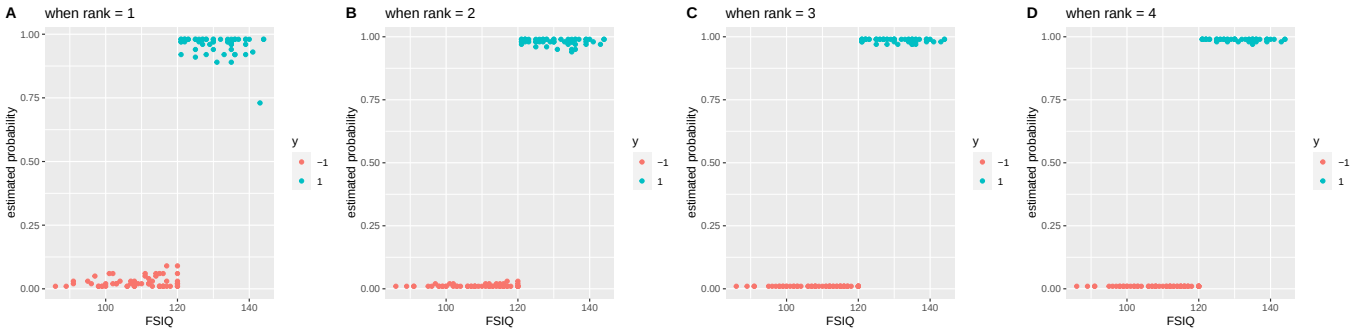


Figure 6: Probability estimation according to the rank $\{1, 2, 3, 4\}$. The x-axis is the true IQ of the subjects and y-axis estimated probability.