Rank estimation

Chanwoo Lee, August 30, 2020

1 Simulation 1

I reduce the size of feature brain connection matrices to 18 by 18 such that all nodes from left and right side match i.e. 9 nodes from each side of brain. First, I made ground truth coefficient $\boldsymbol{B}$ with rank $= 3, 5, 8, 10$ and assign y_i with the following rule

$$y_i = \text{sign}\left(\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle\right), \quad i = 1, \ldots, n. \tag{1}$$

Second, I perform 5-folded cross validation with the same test set and training set with different rank and cost combinations such that $(\text{rank}, \text{cost}) \in \{1, \ldots, 18\} \times \{2, 4, 6, 8, 10\}$. Last, I plot mean accuracy rate from 5 test sets according to rank and cost.

The following figure plot the simulation results. I observed the tendency that as rank increases, cross validation results become similar regardless of cost value. Unfortunately, this simulation shows that we fail to estimate close rank to real rank of ground truth matrix $\boldsymbol{B}$.

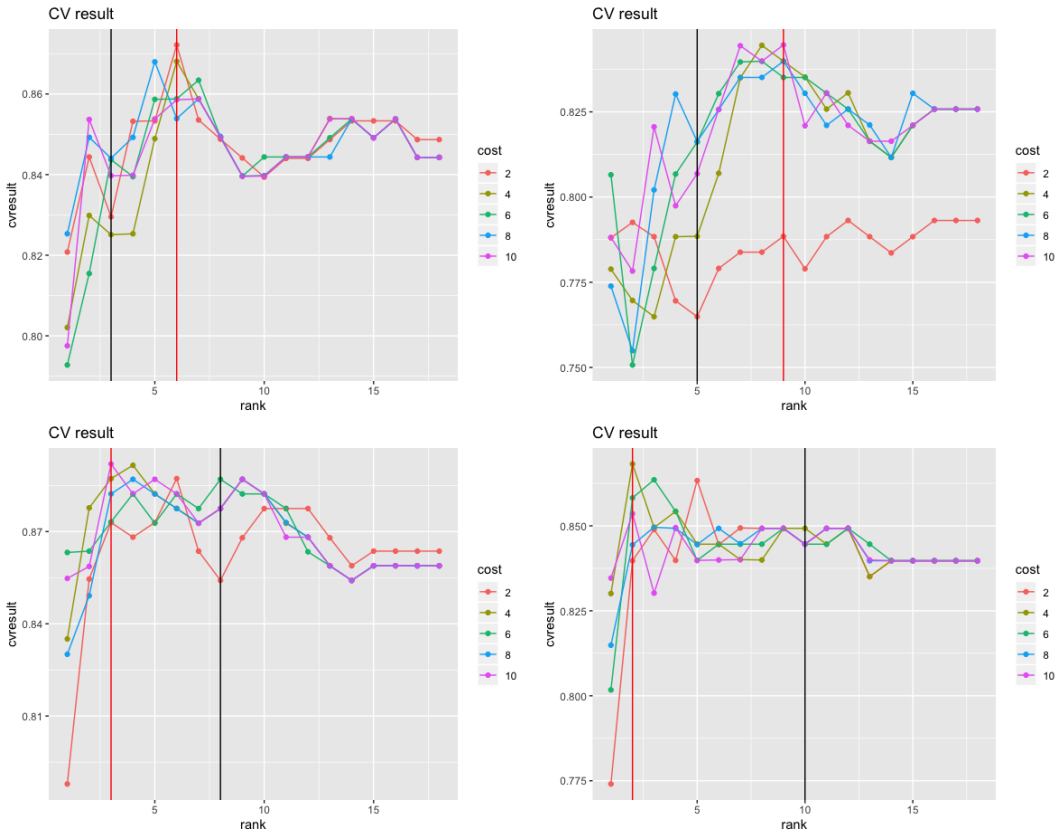

Figure 1: The figures plot mean accuracy rates according to rank and cost values. Black horizontal lines show true rank and red one show the rank which maximizes accuracy rate with certain cost values.

How to explain the inaccurate rank selection?
— Possible statistical reason: (a) small sample size; (b) discrepancy between model loss (hinge loss) vs. evaluation loss (0-1 loss) ==> please verify; s
next page.
(c) low signal. P(Y=1|X) too close to 0.5 for most X's. ==> how to verify?
(d) effect of X?
— Possible computational reason: (d) numerical accuracy in the fitting algorithm. ==> Given a rank, is B well recovered? Try scatter plot of B vs.
estimated B.
plot (Fnorm between estimated B vs. B) vs. rank. Make sure to rescale both B and estimated B to have F-norm 1. Does this lead to correct rank?
Other more reasons?

2 Simulation 2

Since Simulation 1 is not working well, I used centered feature matrices. We define new centered feature matrices as

How to explain?

Why centering of X would have an effect in rank selection?

$$\boldsymbol{X}_i' = \boldsymbol{X}_i - \bar{\boldsymbol{X}}, \quad \text{where } \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

By similar way, I choose ground truth coefficient $\boldsymbol{B}$ with rank $= 3, 5, 8.$ and assign y_i with the rule, $y_i = \text{sign}\left(\langle \boldsymbol{B}, \boldsymbol{X}_i' \rangle\right) = \text{sign}\left(\langle \boldsymbol{B}, \boldsymbol{X}_i - \bar{\boldsymbol{X}} \rangle\right)$ for $i = 1, \ldots, n$. From this data set, I perform 5-folded cross validation by the same way in Simulation 1.

The following figure plot the simulation results. When rank is small (3,5), estimated rank and true rank differ only by 1. However, as rank increase, rank estimation is still bad considering true rank. The tendency of insensitivity to cost value still happens as in Simulation 1 when the rank is large.

Reasonable.

Possible explanation: our accuracy measure is scale-free (0-1 error). If we use hinge loss as the accuracy measure, then we might see nonnegligible effect due to cost ==> please verify

Y-axis: add loss for the training set.
X-axis: rank.

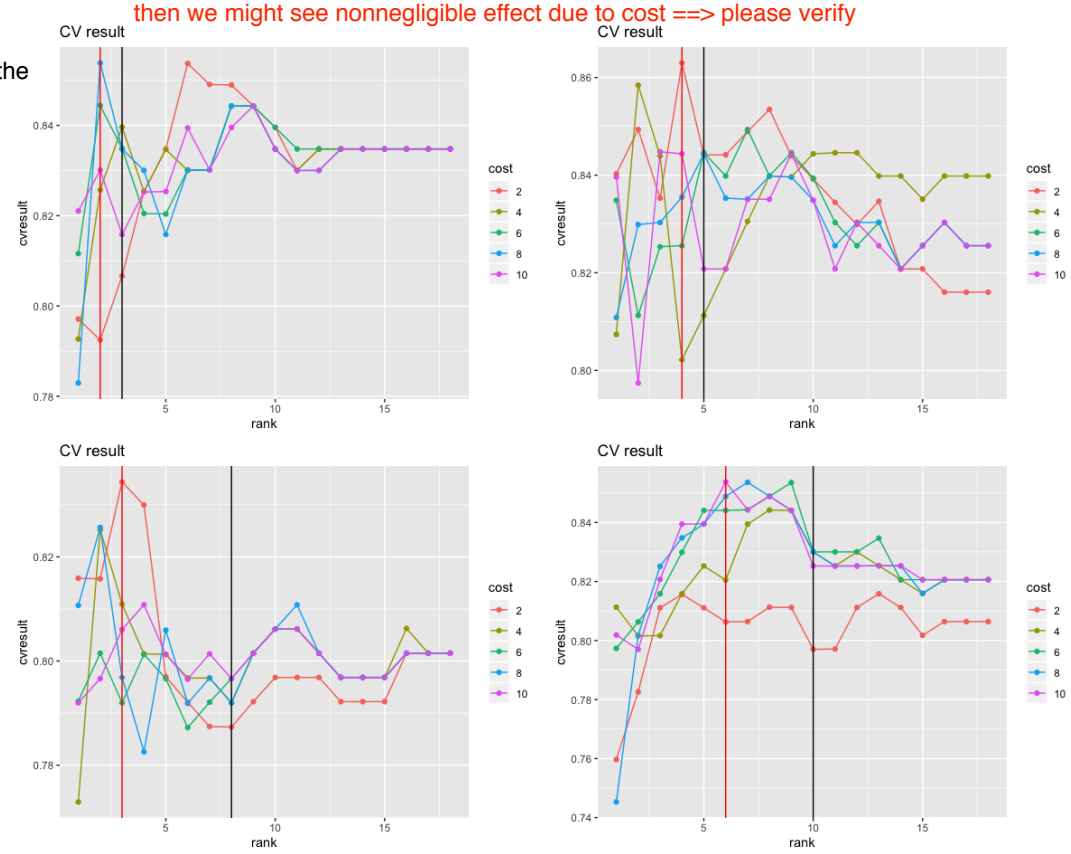

Figure 2: The figures plot mean accuracy rates according to rank and cost values. Black horizontal lines show true rank and red one show the rank which maximizes accuracy rate with certain cost values.

One positive perspective is that simulation 2 performs good when cost $= 6$. The following table shows the estimated rank when cost $= 6$ according to true rank.

	Rank 3	Rank 5	Rank 8	Rank 10
Cost $= 6$	2	7	10	9

I do not understand. How is this way different from what you did earlier?

Table 1: The estimated rank from 5 folded CV when cost $= 6$ according to true rank.

From this perspective, another way to find a good tuning parameter is to have simulation on real sized dataset with ground truth coefficient $\boldsymbol{B}$. From the simulation result, we set a tuning parameter that has the best performance for estimating true rank. To be specific, I randomly generated the coefficient matrix $\boldsymbol{B}$ with rank $r = 3, 5, 8, 10$. By the similar way in Simulation 1, I assigned y_i according to (1). Lastly, I perform 5-folded CV with the same test set and training set with different rank and cost combination such that (rank,cost) $\in \{1, \ldots, 20\} \times \{2, 4, \ldots, 18, 20\}$. My current thought is to choose, my plan is to apply 5-folded CV to estimate the rank which minimizes error rate on test dataset.