BrainData application

Chanwoo Lee, September 6, 2020

1 Data transformation

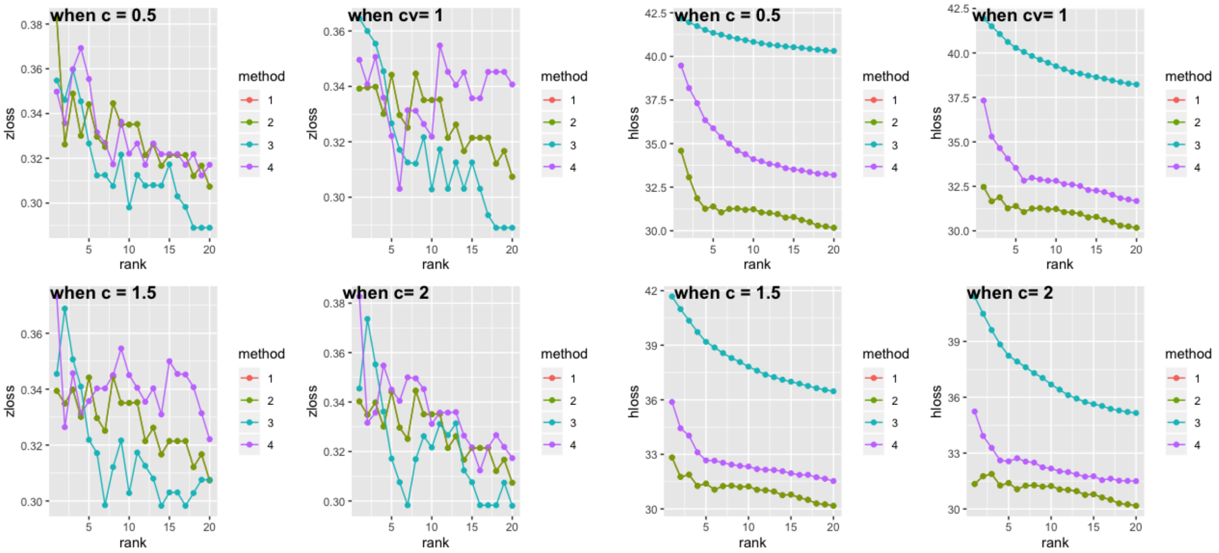
1.1 Simulation setting

I perform 5 folded cross validation based on 4 different data sets. The following shows how data sets are generated from the raw brain data set.

1. Method 1: The raw brain data set $\boldsymbol{X}_i \in \mathbb{R}^{68\times 68}$ for $i = 1,\ldots,212$.
2. Method 2: The centered brain data set such that $\boldsymbol{X}_i^{(2)} = \boldsymbol{X}_i - \bar{\boldsymbol{X}}$ for $i = 1,\ldots,212$, where $\bar{\boldsymbol{X}} = \frac{1}{212}\sum_{i=1}^{212}\boldsymbol{X}_i$.
3. Method 3: The normalized brain data set such that $\boldsymbol{X}_i^{(3)} = \boldsymbol{X}_i/\|\boldsymbol{X}_i\|_F$ for $i = 1,\ldots 212$.
4. Method 4: The normalized and centered brain data set such that $\boldsymbol{X}_i^{(4)} = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})/\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|_F$ for $i = 1,\ldots,212$.

I set all initializations are the same on each simulation by setting the same seed number.

1.2 Simulation result



(a) 0-1 loss evaluation

(b) hinge loss evaluation

Figure 1: The figures plot the loss values according to given rank and cost in $\{1,\ldots,20\} \times \{0.5, 1, 1.5, 2\}$

I found one interesting phenomenon. Method 1 and Method 2 have the exactly same output results. One reason I did not observe this phenomenon in previous experiment is that I did not use the same true coefficient matrix $\boldsymbol{B}$. I checked that with the same true coefficient matrix $\boldsymbol{B}$ and same set.seed number, the outputs from Method 1 and Method 2 are the same. Figure 1 shows that Method 1 works best in hinge loss evaluation and Method 3 performs best in 0-1 loss. From this simulation I decided to use Method 1 because if we use Method 3, we might lose direct interpretation of coefficient matrix $\boldsymbol{B}$. To be specific, without normalization each entry of $\boldsymbol{B}$ represents how important the brain connection is, which can be applied to all individual with the same magnitude. However, the weight of importance is changed to each individual if we use normalized $\boldsymbol{B}$.

2 Cross Validation results.

First simulation is to find overall range of cost value. I perform cross validation with cost values in $\{0.01, 0.1, 1, 10, 100\}$. I set the 3 multiple initializations to estimate the classifiers. We can see that cost values greater than 0.1 converges to the same output with cost $= 0.1$ as the rank size increases.

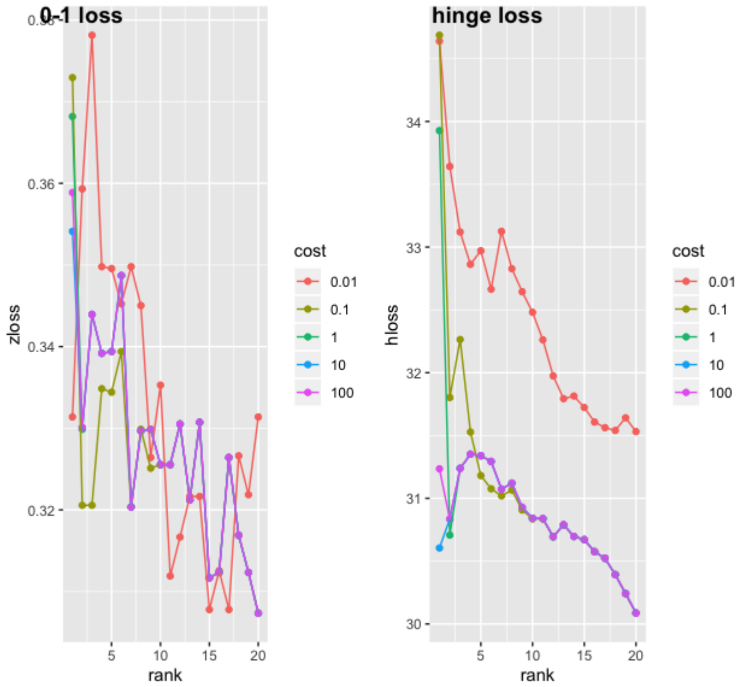


Figure 2: Cross validation results. The left figure shows 0-1 loss evaluation according to rank and cost combinations while the right one shows hinge loss evaluation.

Therefore, I changed range of cost value to $\{0.005, 0.01, 0.05, 0.1, 0.15, 1\}$ and rank to $\{1, \ldots, 30\}$. To have stable result, I changed the repetition from 3 to 10 and make each trial have the same initial point across all combinations of (rank, cost) for fair comparison. The following plot shows the cross validation result The smallest 0-1 loss is 0.3030303 with (rank,cost) $\in \{(23, 0.05), (23, 0.1), (23, 0.15), (23, 1)\}$. When the rank size is greater than 20, cost values ranging from 0.05 to 1 have the exactly the same performance and hinge loss is monotonically decreasing to the rank size. If we focus on the ranks ranging from 0 to 10, there are local minimums with respect to both 0-1 loss and hinge loss so that we can pick for estimated rank. Currently, I am running a code for ranks ranging from 30 to 60 to make sure rank 23 is global minimum when 0-1 loss is used.

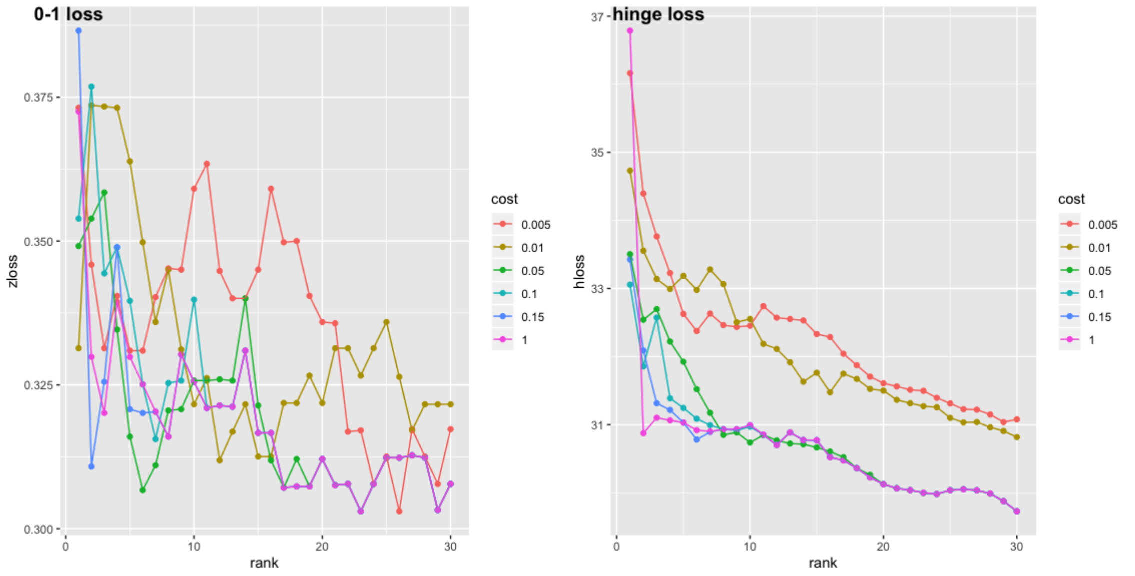


Figure 3: Cross validation results. The left figure shows 0-1 loss evaluation according to rank and cost combinations while the right one shows hinge loss evaluation.