VSPLOT data analysis and comparsion

Chanwoo Lee, October 19, 2020

1 VSPLOT brain data log-likelihood

Table 1 shows the top 10 best log-likelihood on test datasets in cross validation. I compared old SMMK and ADMM versus new SMMK and ADMM algorithm based on log-likelihood on the full dataset. We can see new SMMK algorithm works worse than previous one while ADMM algorithm improved.

rank	sparsity	log_L_test	log_L_train	old SMMK	old ADMM	new SMMK	new ADMM
1	65	-29.08	-107.69	-124.84	-138.22	-145.82	-131.48
1	66	-29.24	-112.63	-135.72	-143.08	-146.99	-133.11
1	68	-29.40	-117.59	-141.70	-146.91	-146.99	-146.99
1	67	-29.49	-115.30	-142.74	-146.66	-146.99	-144.85
1	64	-29.59	-102.80	-117.96	-128.77	-146.99	-124.47
2	67	-30.30	-112.06	-135.39	-129.66	-143.26	-129.35
1	62	-30.31	-96.85	-110.71	-132.70	-145.06	-127.90
1	63	-30.62	-100.48	-117.24	-134.37	-145.68	-131.50
2	66	-30.87	-107.67	-136.79	-132.96	-142.32	-132.28
2	64	-31.39	-102.15	-119.84	-116.05	-144.02	-126.43

Table 1: Top 10 combinations of rank and sparisty in log-likelihood of test datsets. The third and fourth columns represents averaged log-likelihood of test and training datasets in 5-folded cross validation. The last four columns shows log-likelihood values on the full datasets according to each algorithm used.

Figure 1 shows the log-likelihood trend according to different algorithms (latest one).

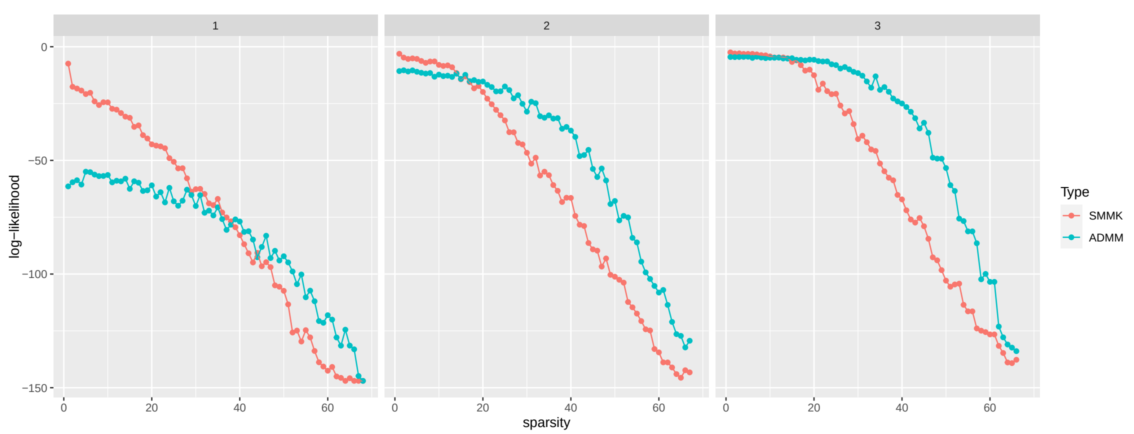


Figure 1: The figure shows the log-likelihood values according to sparsity given rank = 1,2, and 3. The colors represent the algorithm type used.

2 Comparison to alternative method

I want to check whether the result from hyper-parameter (rank, sparsity) = (1,65) is good enough compared to alternative methods. For an alternative method to compare with ours, I used logistic-lasso regression method. I briefly introduce this method. First, I obtained vector x_i from the adjacency matrices $\boldsymbol{X}_i$ only saving the lower triangular elements of $\boldsymbol{X}_i$.

$$x_i = (\boldsymbol{X}_{i[21]}, \boldsymbol{X}_{i[31]}, \ldots, \boldsymbol{X}_{i[68(67)]})^T.$$

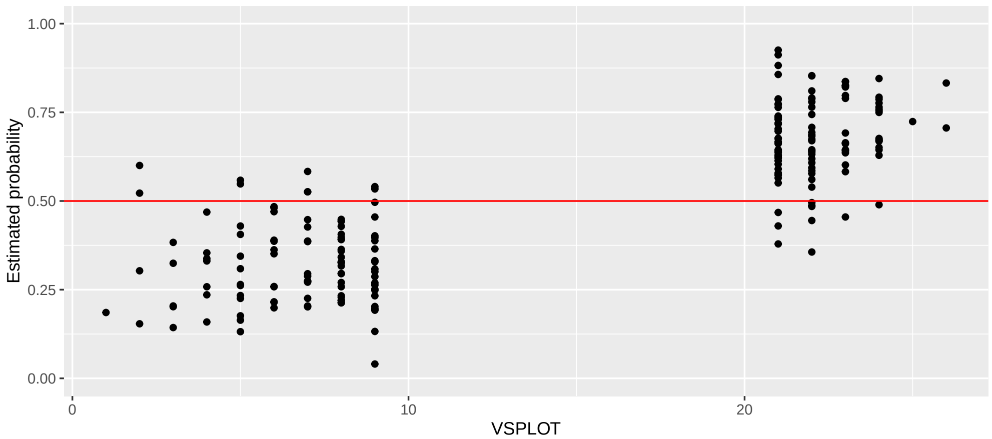
I set the logistic model as

$$p(y_i = 1|\boldsymbol{x}_i) = \frac{e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i}}.$$

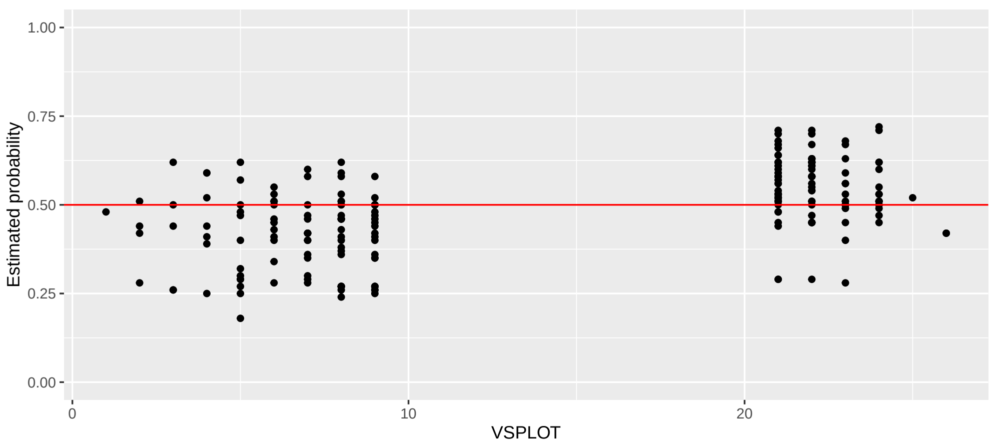
I use the variable selection using lasso method from the following loss function

$$-\sum_{y_i=1} \log\left(p(y_i = 1|\boldsymbol{x}_i)\right) - \sum_{y_i=-1} \log\left(1 - p(y_i = 1|\boldsymbol{x}_i)\right) + \lambda \sum_i |\beta_i|.$$

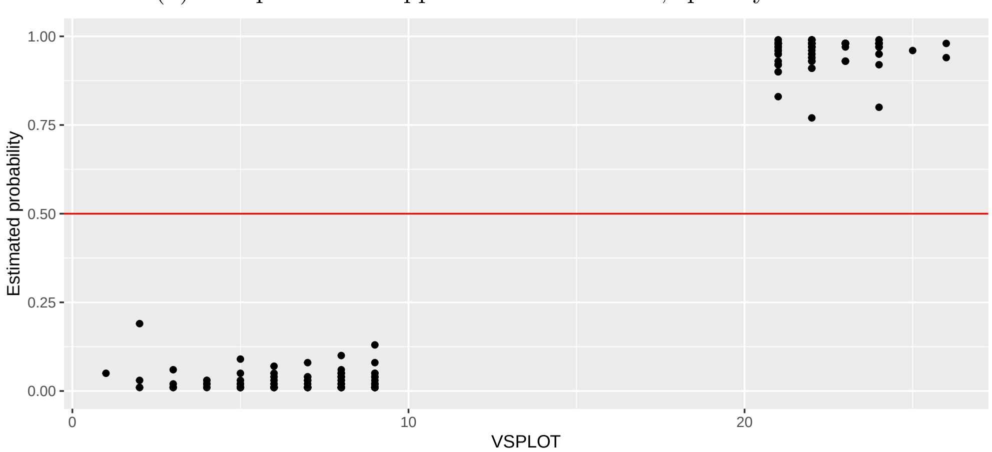
An optimal λ is obtained from a grid search. I used the function `cv.glmnet` in `glmnet` r-package for λ. Based on the dataset $\{(x_i, y_i)\}_{i=1}^{212}$, I calculated the estimated probabilities $\mathbb{P}(y_i = 1|x_i)$. Figure 2 (a) shows the logistic-lasso regression result. Our method is shown in Figure 2 (b) and (c), where (b) uses (rank,sparsity) = (1,65) while (c) uses rank = 1 without sparsity. It seems to me that (c),(a), and (b) performs better in order. One interesting result I found is when I perform cross validation on this alternative model, the averaged log-likelihood on test datasets is -26.84, which is greater than the log-likelihood based on our model. Furthermore, it has -79.82 as averaged log-likelihood values on training datsets. I performed the simple simulation I did in previous note where $\boldsymbol{X}_i \in \mathbb{R}^{5\times 5}$ and $y_i = \frac{e^{\text{sign}(\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle)}}{1+e^{\text{sign}(\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle)}}$, I have checked that the logistic-lasso gives me random guess.



(a) Logistic-lasso.



(b) Non-parametric approach with rank =1, sparsity = 65.



(c) Non-parametric approach with rank =1 without sparsity.

Figure 2: Averaged log-likelihood value on test datasets according to the types of cross validation.