Algorithm sanity check

Chanwoo Lee, October 29, 2020

1 SMMK improvement

I obtained parts of cross validation results from new updated algorithm. Some jobs stopped because of out of memory: Modified algorithm takes more than 14 GB to run properly. So I plotted available combinations. For comparison, I plot the cv result before modification (Figure 1). Based on cv result on test datasets, we can check the new algorithm improved to find points that minimizes the loss function. It seems that ADMM performs better under the sparse coefficient regime while SMMK works better under the dense regime. This results are based on the smoothing parameter H = 99. I requested jobs on CHTC server with different smoothing parameter H = 49 to check if there is improvement.

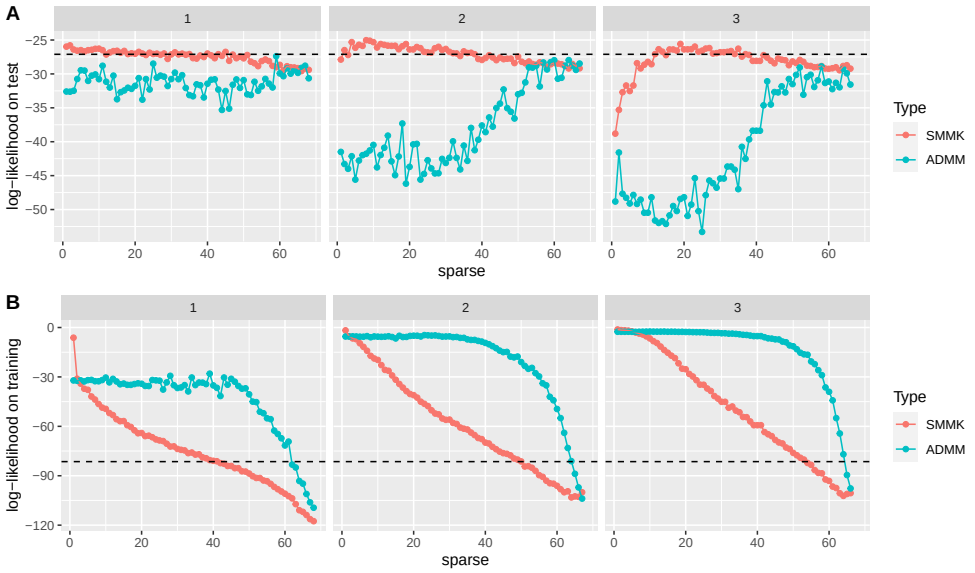
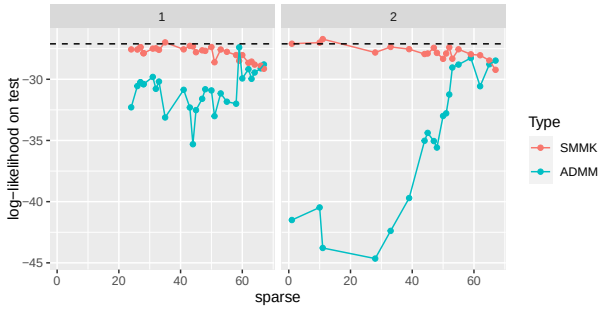
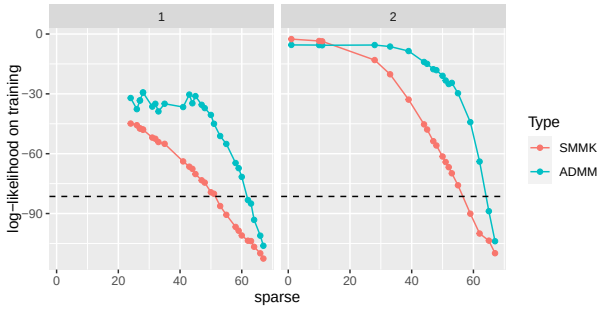


Figure 1: CV results before modification



(a) CV result on test datasets



(b) CV result on training datasets

Figure 2: CV results after modification: (a) corresponds to Figure 1-A while (b) to Figure 1-B. The dotted lines are the averaged log-likelihood values from lasso logistic regression.

2 Sanity check from a simulation

I generated 100 dataset $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^{100}$ where $\boldsymbol{X}_i \in \{0,1\}^{10\times 10}$ are binary feature matrices. The corresponding label is decided by

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

where $\boldsymbol{B} \in \mathbb{R}^{10\times 10}$ is the coefficient matrix whose rank is 5 and sparsity is 3 such that 2nd, 4th, and 9th rows and columns are 0. Figure 3 shows the true probability and labels of the dataset.

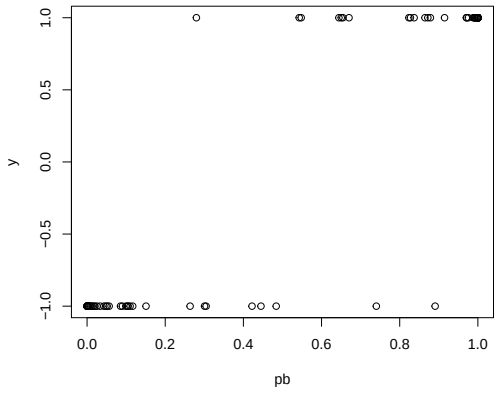


Figure 3: Ground truth of the probabilities and labels.

2.1 Classification

First, I check whether the algorithms successfully ground truth matrix $\boldsymbol{B}$ given the true (rank, sparsity) = (5,3). Figure 4 plots the normalized true $\boldsymbol{B}$ versus the normalized estimated $\hat{\boldsymbol{B}}$ according to methods: SMMK with option 1, SMMK with option 2, ADMM, and lasso logistic. All our methods work quite good to estimate the true $\boldsymbol{B}$. In addition, SMMK method finds right sparse rows and columns (2,4,9) while ADMM have (4,9,10) rows and columns being zero. In addition, I have checked that SMMK has smaller objective value than ADMM when I calculated

$$\text{obj}(\boldsymbol{B}, y, p) = \frac{1}{2}\|\boldsymbol{B}\|_F^2 + (1-p)\sum_{y_i=1}(1 - y_i(\langle \boldsymbol{B}, \boldsymbol{X}_i\rangle + b))_+ + p\sum_{y_i=-1}(1 - y_i(\langle \boldsymbol{B}, \boldsymbol{X}_i\rangle + b))_+ .$$

2.2 Probability estimation

Next, I checked whether the algorithms works good to estimate the ground truth probabilities based on the dataset. Here I set the smoothing parameter $H = 49$. I perform cross validation to find the ground truth combinations of (rank, sparsity) from the result on test datsets. Since it takes a lot of time to see all combinations, I only inspected the rank 1, 3, 5, and 7. It turns out that SMMK shows the best performance on (rank, sparsity) = (5,2) while ADMM has (rank, sparsity) = (5,4). Considering the fact that the ground truth combination is (5,3), both methods have reasonable result. In addition, Figure 5 plots the averaged log-likelihood values across different rank and

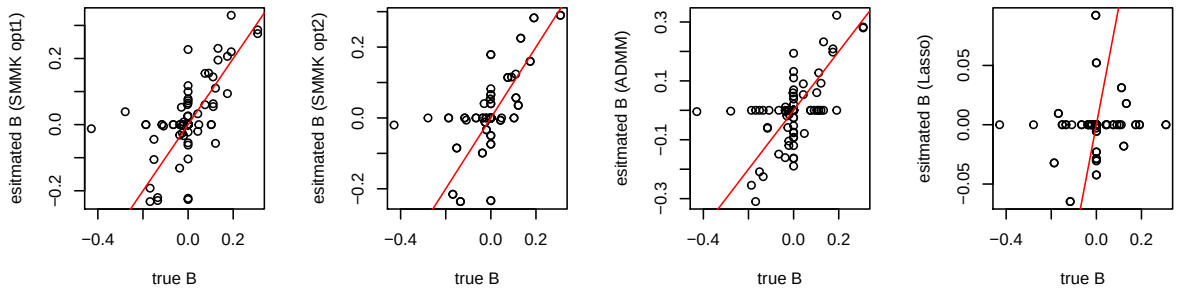
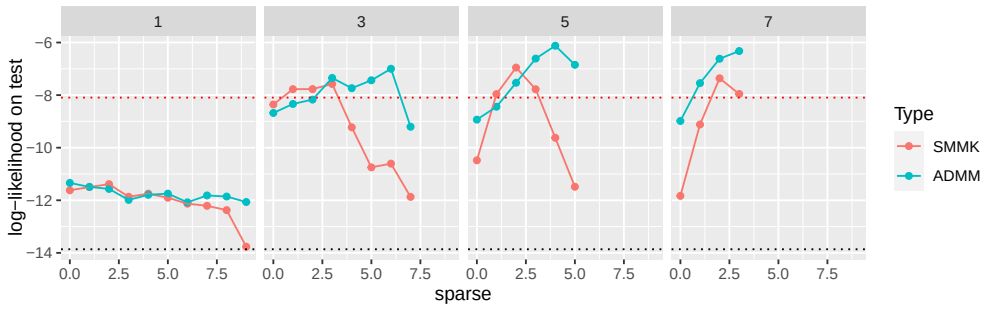


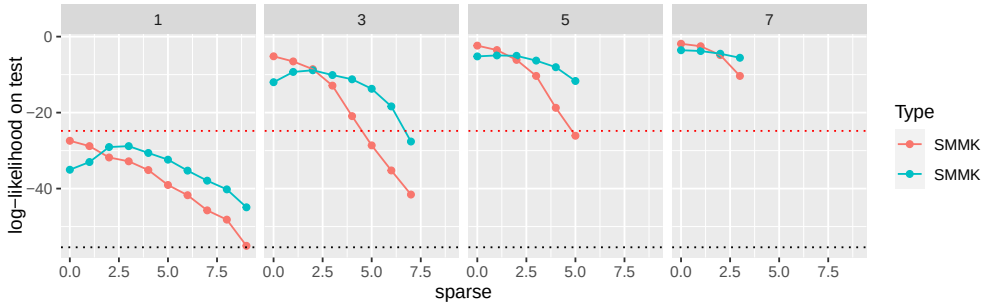
Figure 4: First three figures are based on large margin classification model (SMMK option 1, SMMK option 2, ADMM in order) while the last figure is from lasso logistic model. I used hyper-parameters (rank, sparsity, weight) = (5, 3, 0.5) in the large margin model.

sparsity on test and training datasets. CV result on test datasets verifies that ADMM fits better on sparse regime while SMMK fits better on dense regime. Log-likelihood on test dataset from ADMM shows decreasing trend when sparsity is small, which is similar to brain VSPLOT dataset case (Figure 1). Overall, ADMM seems to perform better than SMMK. This simulation shows that the trend of each algorithm according to sparsity are the same in real dataset. But the decreasing magnitude of log-likelihood from ADMM when sparsity is small cannot be explained well. I have double checked any possible bugs in the cross validation algorithm but have not found critical bugs. Hopefully, new cross validation result I am waiting for based on ADMM with $H = 49$ might give us improved result similar to the result on the simulation.

Figure 6 plots the true probabilities versus estimated one at the best rank and sparsity combination of each method ((5,2) for SMMK and (5,4) for ADMM). ADMM based probability estimation gives us more informative result than SMMK.



(a) CV result on test datasets



(b) CV result on training datasets

Figure 5: CV results after modification: (a) corresponds to Figure 1-A while (b) to Figure 1-B. Red dotted lines shows the performance of lasso logistic method and black dotted lines are the performance of random guess.

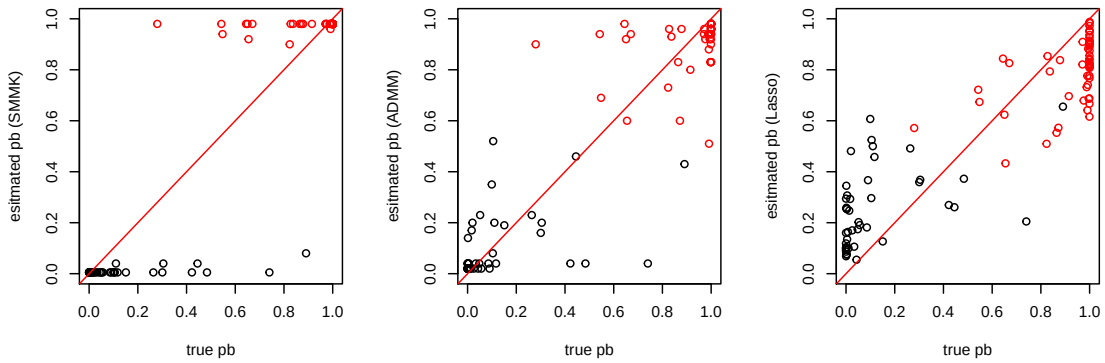


Figure 6: Estimated probabilities versus ground truth probabilities. The first figure is from SMMK algorithm with (rank, sparsity) = (5,2). The second figure is from ADMM with (rank, sparsity) = (5,4). The last figure is from logistic lasso approach.