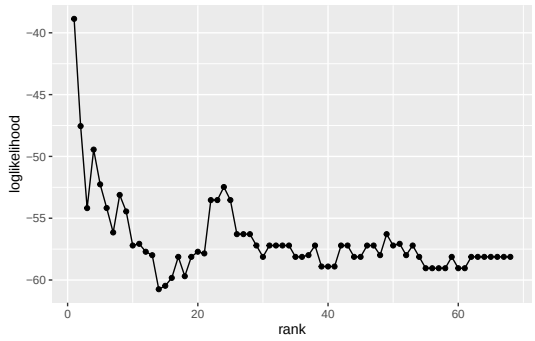
Sanity check of loglikelihood cross validation

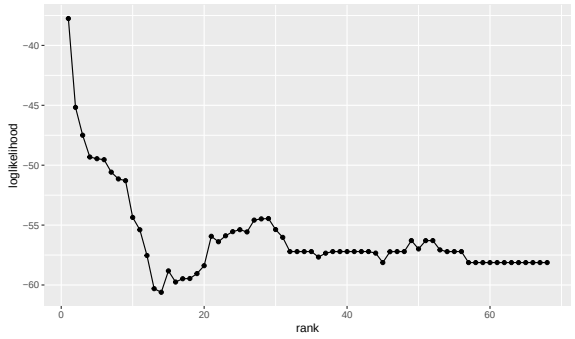
Chanwoo Lee, October 11, 2020

1 SMMK versus ADMM in brain IQ data analysis

I obtain averaged log-likelihood on test datasets in cross validation. Figure 1 shows that two results based on SMMK and ADMM are not much different.



(a) Cross validation result based on ADMM algorithm



(b) Cross validation result based on SMMK algorithm

Figure 1: Averaged log-likelihood value on test datasets when no sparsity strucure is imposed.

2 Simulations based on two different settings

I generated the datasets whose feature matrices $\boldsymbol{X}_i \in \mathbb{R}^{5\times 5}$ are symmetric binary matrices. From given $\boldsymbol{B} \in \mathbb{R}^{5\times 5}$ whose true rank is 3, I assigned the label responses y_i in two different ways.

Sim 1 (without noise case)

$$y_i^{(1)} = \text{sign}\left(\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle\right).$$

Sim 2 (with noise case):

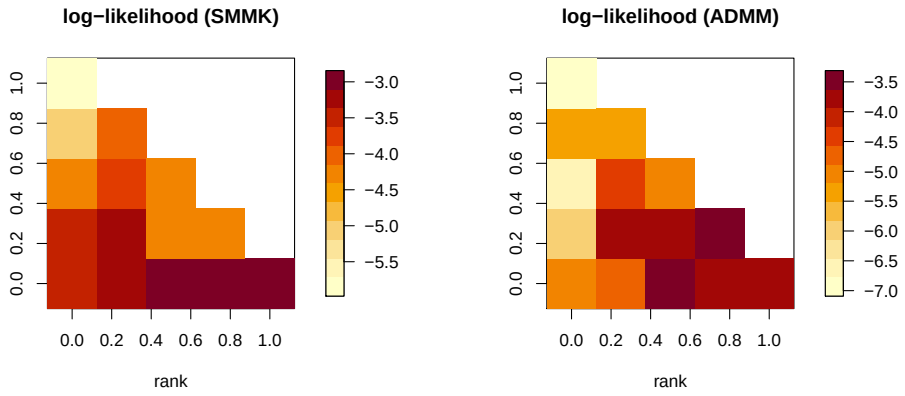
$$y_i^{(2)} \overset{\text{ind}}{\sim} \text{Ber}\left(\text{logistic}(\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle)\right).$$

My conjecture of cross validation picking (rank, sparsity) = (1,68) as top 3 from IQ brain datset is that we do not have enough sample size. To be specific, brain data has matrix predictors whose size is 68 by 68 but the number of sample size is 114. So I simulated two cases: Large sample case ($n = 50$) and small sample case ($n = 10$). Then I check how cross validation results change.

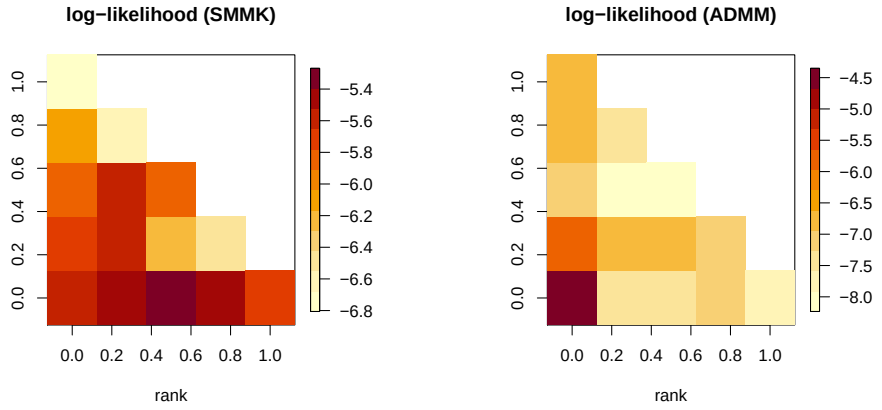
2.1 Large sample case

I generated 50 observations $\{(\boldsymbol{X}_i, y_i^{(1)})\}_{i=1}^{50}$ and $\{(\boldsymbol{X}_i, y_i^{(2)})\}_{i=1}^{50}$ where $y_i^{(1)}$ is generated from Sim 1 while $y_i^{(2)}$ from Sim 2.

Figure 2 shows the averaged log-likelihood across the combination (rank, sparsity). In Sim 1, the best (rank, sparsity) is (4,1) based on SMMk while (3,1) on ADMM. Considering that the true rank



(a) Cross validation result based on Sim 1 (without noiese).



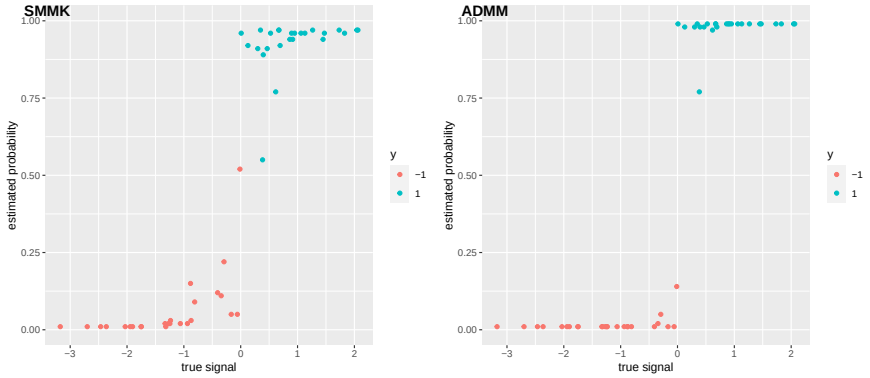
(b) Cross validation result based on Sim 2 (with noise).

Figure 2: Averaged log-likelihood value on test datasets in large sample case (n = 50).

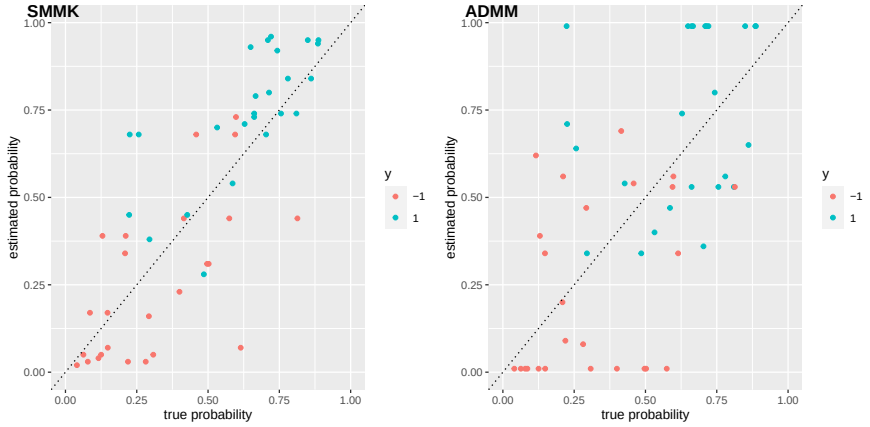
is 3 and sparsity is 1, both cross-validation based hyper-parameter selection works well. In Sim 2, the best (rank, sparsity) is (3,1) based on SMMK while (1,1) on ADMM. This shows ADMM method is less stable than SMMK to pick the best hyper-parameters in large sample case.

Figure 3 shows the probability estimation results in Sim1 and Sim2 at the best (rank, sparsity) combination based on Figure 2. Though the result from ADMM method is a little bit deviate from true probability in Sim 2, All estimation results look reasonable.

The simulation result shows that both SMMK and ADMM algorithm work well in probability estimation when the sample size is large enough.



(a) Probability estimation result based on Sim 1 (without noise)/



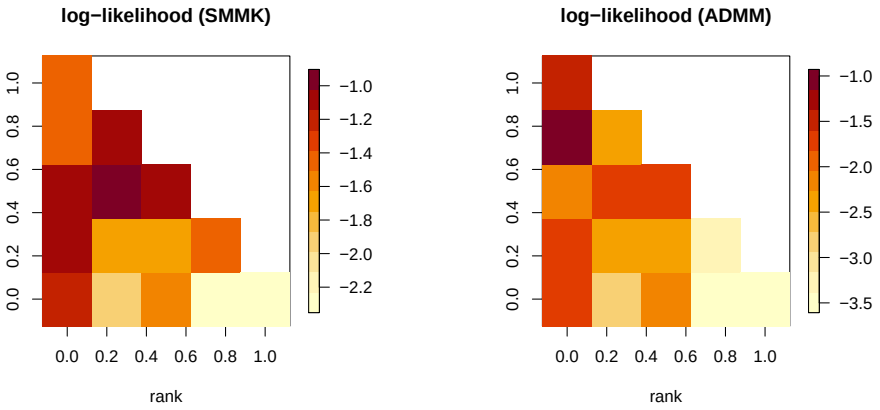
(b) Cross validation result based on Sim 2 (with noise).

Figure 3: Probability estimation results setting the best (rank, sparsity) in each algorithm that maximize log-likelihood on test datset. the rank and sparsity are set as (4,1) for SMMK and (3,1) for ADMM in Sim 1 while (3,1) for SMMK and (1,1) for ADMM in Sim 2. The number of sample size is 50.

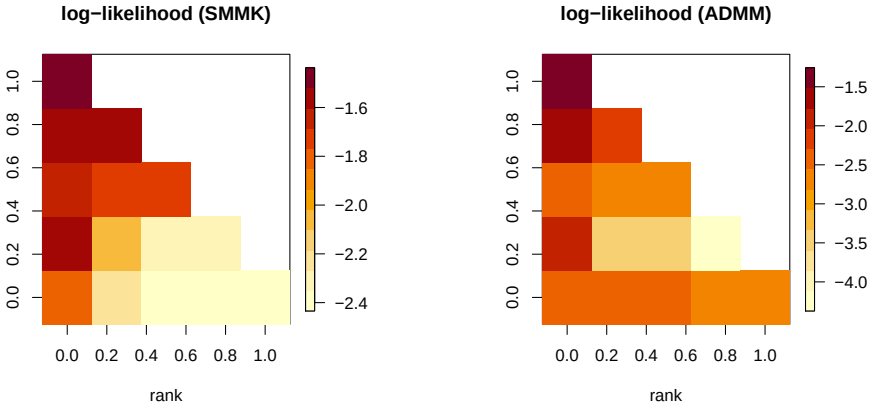
2.2 Small sample case

I generated 10 observations $\{(\boldsymbol{X}_i, y_i^{(1)})\}_{i=1}^{10}$ and $\{(\boldsymbol{X}_i, y_i^{(2)})\}_{i=1}^{10}$ where $y_i^{(1)}$ is generated from Sim 1 while $y_i^{(2)}$ from Sim 2. This sample size is somewhat similar to the case of brain IQ dataset. Figure 4 shows the quite similar trend that we have in brain IQ data. In both Sim 1 and Sim 2 cases, we can see that (rank, sparsity) combinations which has less the number of parameter are favorable to be picked as the best regardless of SMMK and ADMM algorithms. To be specific, (2,3) is picked when SMMK is used and (1,4) when ADMM is used in Sim 1 (without noise). In Sim 2 (with noise), the best combination is (1,4) in both SMMK and ADMM. Since true combination is (3,1), averaged log-likelihood on test dataset does not offer good hyper-parameters.

Based on the cross validation result, I estimated probability according to the best combination for each algorithm and simulation. In Sim 1, we can not obtain meaningful probability estimation because the dataset is completely separable thought the number of sample size is small. Unlike the large sample size case, we do not have estimated probabilities near 0.5 around the threshold 0. In simulation 2, we fail to obtain good probability estimation with (rank, sparsity) combinations from log-likelihood based cross validation. One interesting thing to notice is the result based on ADMM



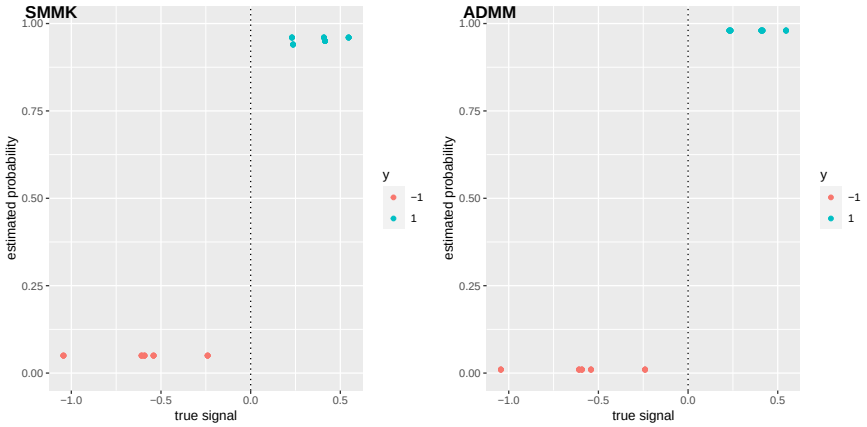
(a) Cross validation result based on Sim 1 (without noise).



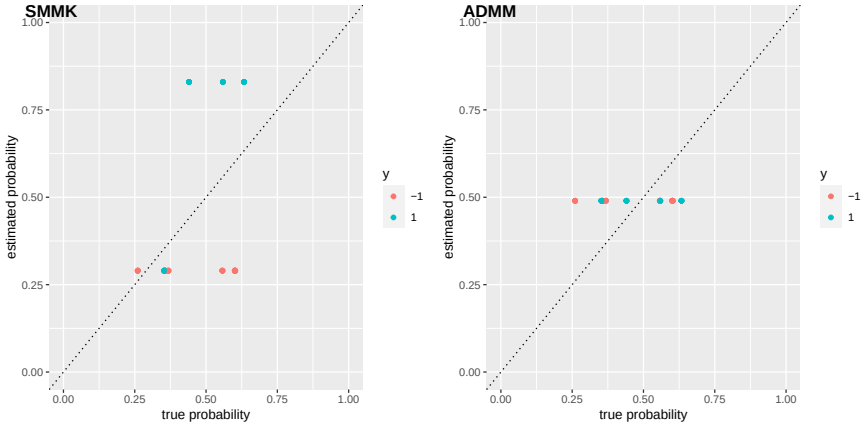
(b) Cross validation result based on Sim 2 (with noise).

Figure 4: Averaged log-likelihood value on test datasets in small sample case (n = 10).

in Sim 2 shows the exactly the same result from the brain IQ data.



(a) Probability estimation result based on Sim 1 (without noise)



(b) Cross validation result based on Sim 2 (with noise).

Figure 5: Probability estimation results setting the best (rank, sparsity) in each algorithm that maximize log-likelihood on test datset. the rank and sparsity are set as (2,3) for SMMK and (1,4) for ADMM in Sim 1. (1,4) is used for both SMMK and ADMM in Sim 2. The number of sample size is 50.

3 New way to pick the hyper parameters

When the sample size is small, log-likelihood on test datasets does not give us good hyper-parameters. One of the reason is that the combinations of (rank, sparisty) which has high log-likelihood values fail to learn training dataset well. For example, the combination (1,5) give us the the maximum log-likelihoods -1.47 for SMMK and -1.38 for ADMM on test datasets while giving us the minimum log-likelihoods -3.86 for SMMK and -5.54 for ADMM. You can check detailed trends in Figure 6 (a) and (b). Similar phenomenon happens in Sim 1. Since the number of sample size is too small, consideration on only test dataset (2 data points for each cross validation in simulation) cannot reflect the model performance well. Therefore, I suggest to consider log-likelihood values on both test and training datsets. To be specific, log-likelihood on test data or training data set is calculated as

$$l(test) = \frac{1}{5}\sum_{k=1}^{5}\left(\sum_{i\in\{y_i=1, i\in \text{test}_k\}} log(p_i) \sum_{i\in\{y_i=-1, i\in \text{test}_k\}} log(1-p_i)\right),$$

$$l(train) = \frac{1}{5}\sum_{k=1}^{5}\left(\sum_{i\in\{y_i=1, i\in \text{train}_k\}} log(p_i) \sum_{i\in\{y_i=-1, i\in \text{train}_k\}} log(1-p_i)\right),$$

where test_k is a test index set and train_k is a training index set at k-th cross validation. We consider the log-likelihood on both test and training datasets by adding them together.

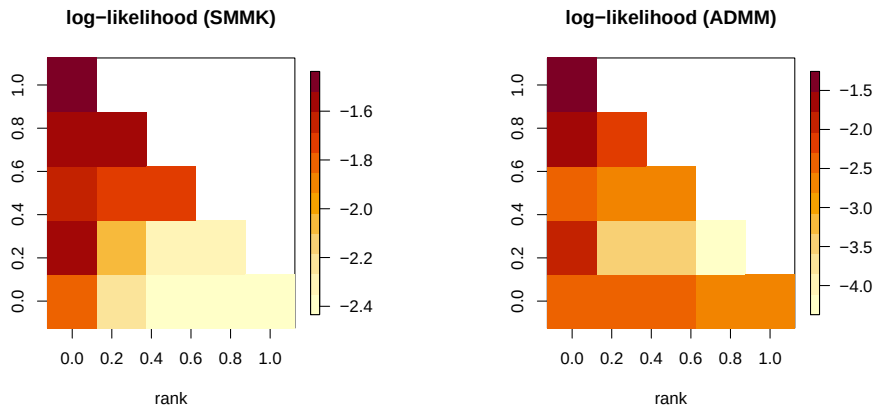
$$l(data) = l(test) + l(train).$$

By picking the hyper-parameter that maixmizes $l(data)$, we can find good one which prevent overfitting but has moderate fitting on training dataset. The following table is the summary of estimated2, hyper-parameter when we consider only test, only training, and both in Sim 1 and Sim 2. Table 1 shows much improved hyper-parameter estimation when both traning and test dataset are considered.

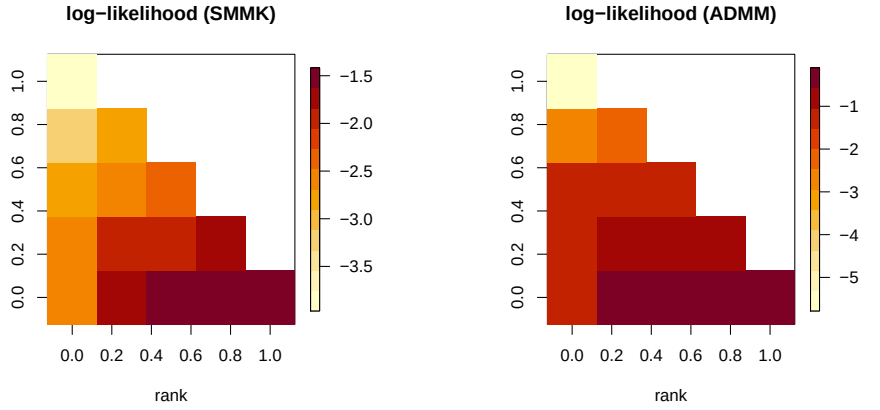
	Based on test datasets	Based on training datasets	Based on both
Sim 1	(2,3) / (1,4)	(5,1) / (5,1)	(3,1) / (3,1)
Sim 2	(1,5) / (1,5)	(4,1) / (5,1)	(4,1) / (3,1)

Table 1: Best hyper-parameter combinations (rank, sparsity) based on each method. First combination is when SMMK is used while the second one is when ADMM is used. True rank and sparsity combination is (3,1).

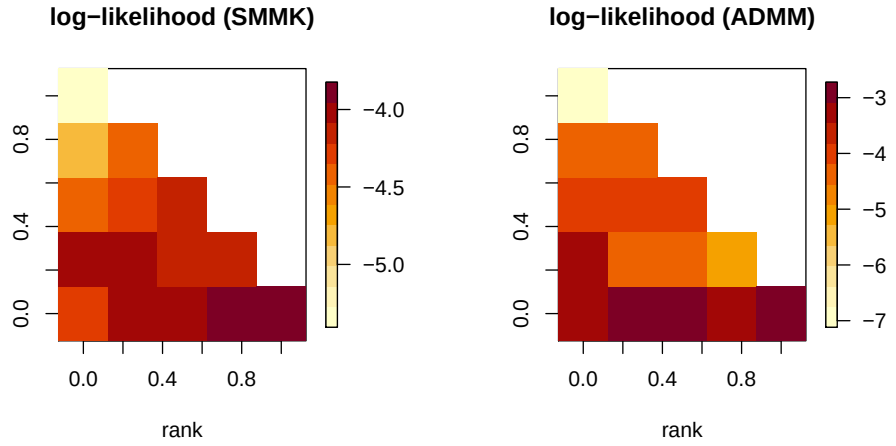
Figure 7 plots the estimated probability versus true probability in Sim 2. As one can see, hyper-parameters decided by the suggested criteria show more informative probability estimation than previous estimation results.



(a) Averaged log-likelihood on test datasets in Sim 2.

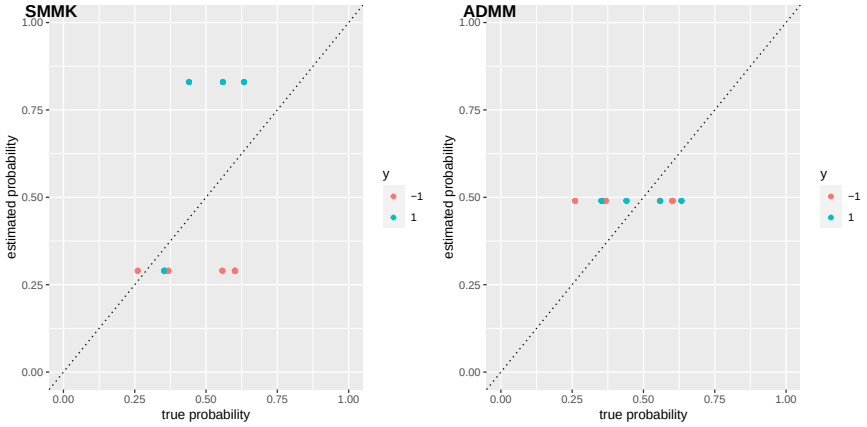


(b) Averaged log-likelihood on train datasets in Sim 2.

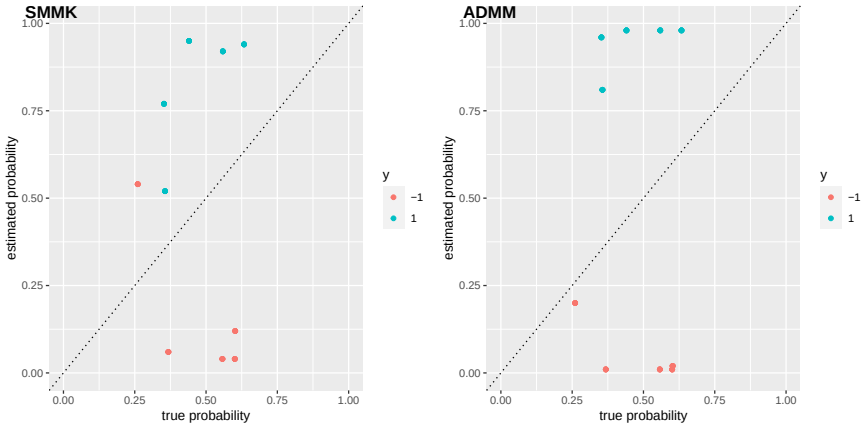


(c) Averaged log-likelihood across test and train datasets in Sim 2.

Figure 6: Averaged log-likelihood value on either test datasets or training datsets. Figure (a) considers only test datasets while Figure (b) considers training datasets. Figure (a) and (b) show clear different trends: less complex model works better on test datasets while more complex model works better on training datsets. Figure (c) considers both test and training datasets.



(a) Probability estimation result in Sim 2. The hyper-parameters are set based on log-likelihood values on test datsets.



(b) Probability estimation result in Sim 2. The hyper-parameters are set based on log-likelihood values on both test and training datsets.

Figure 7: Probability estimation result in Sim 2. In Figure (a), (rank, sparsity) = (1,5) is set in both SMMK and ADMM algorithms. In Figure (b), (rank, sparsit) is set to be (4,1) in SMMK and (3,1) in ADMM.