Cross validation on small sample size and BIC

Chanwoo Lee, October 15, 2020

1 Cross validation ratio

I generated 10 samples from the following rule, whose feature matrices $\boldsymbol{X}_i \in \mathbb{R}^{5\times 5}$

$$y_i \overset{\text{ind}}{\sim} \text{Ber}\left(\text{logistic}(k * \langle \boldsymbol{B}, \boldsymbol{X}_i \rangle)\right), \tag{1}$$

where $\boldsymbol{B}$ is rank 3 coefficient matrix.

If k is increased enough, then (1) rule acts like

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

which I already simulated in the previous note. The following table is true probability and corresponding labels when $k = 3$ in (1).

prob	0.04	0.14	0.14	0.16	0.33	0.67	0.67	0.77	0.78	0.84
label	-1	-1	-1	1	-1	-1.00	1.00	1.00	1.00	1.00

When we set the true hyper-parameters (rank,sparsity) = (3,1), two algorithms SMMK and ADMM estimate the conditional probability $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i)$ for $i \in [10]$ as in Figure 1. It seems that SMMK algorithm estimate the probabilities quite well when true hyper-parmeters are used while ADMM algorithm tends to estimate probability similar to labels of the dataset.

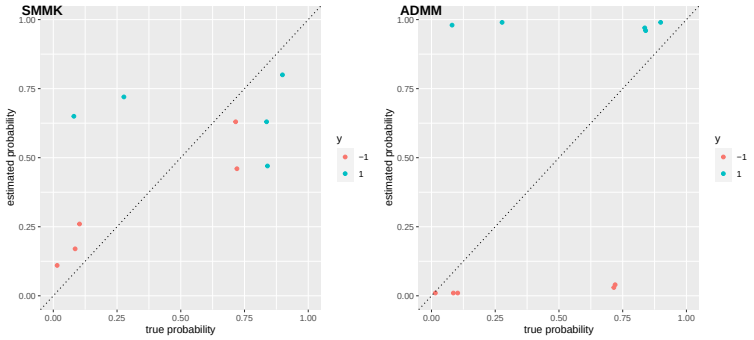


Figure 1: Estimated conditional probabilities when true hyper-parameters are set

I performed 2-folded, 5-folded, and 10-folded (LOOCV). It turns out that the combination (rank, sparsity) = (1,5) has the maximum averaged log-likelihood on test datasets across all cross validation while the combination (rank, sparsity) = (5,1) has the maximum values on test datasets. The following figure shows the averaged log-likelihoods on test datsets according to the types of cross validation.

There have been many papers which study whether cross validation method is valid for small sample [1, 2, 4, 5] Main point of those paper is that regular k-folded cross validation result displays excessive variance and is biased, which makes individual estimates unreliable for small samples. None of the papers suggests good solution we can use on our dataset. For example, some paper

[2, 1] focuses on showing invalidity of cross validation, another paper [5] suggests method how to increase data set based on data sharing and pooling, and the other papers [4] tackle the comparison of models.

2 BIC

From the invalidity of cross validation for small sample size, I perform BIC

$$BIC(r, sparse) = -2\left(\sum_{y_i=1}\log\left(\hat{\mathbb{P}}(y_i = 1|\boldsymbol{X}_i, r, sparse)\right) + \sum_{y_i=-1}\log\left(1 - \hat{\mathbb{P}}(y_i = 1|\boldsymbol{X}_i, r, sparse)\right)\right)$$

$$+ p_e(r, sparse)\log(n),$$

where $p_e(r, sparse) \overset{\text{def}}{=} (d_1 + d_2 + 2 - 2sparse - r)r$ is the effective number of parameters. However, I obtained that (rank, sparsity) = (1,5) is the best based on BIC values. which is the same result from cross validation. Regular BIC method turns out to be defective when the dimensionality of covariates is not fixed which is common in high-dimensional regime. There are several papers to modify BIC to have consistent model estimation in high-dimensional regime [3, 6]. However, the modified BIC penalize model complexity term more, which makes the rank from modified BIC small. Since the simulation setting has too small sample size, I want to check if the real dataset has the same result as cross-validation had. I am running the code for brain dataset whose sample size 212 to see BIC based hyper-parameter selection result.

3 Preliminary result from VSPLOT brain dataset

I performed BIC result when the rank = 1 with different sparsity to set the hyper parameter. Figure 4 plots the BIC value across different sparsities. BIC values has the smallest when the rank = 1 and sparsity = 61. So I set the hyperparameter as (rank,sparsity) = (1,61), estimated conditional probabilities of data points, and compare that probabilities and actual VSPLOT of each individual. Figure 5 shows that estimated probability has linear tendency with respect to VSPLOT score.

References

[1] Daniel Berrar, Ian Bradbury, and Werner Dubitzky. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics*, 22(10):1245–1250, 2006.

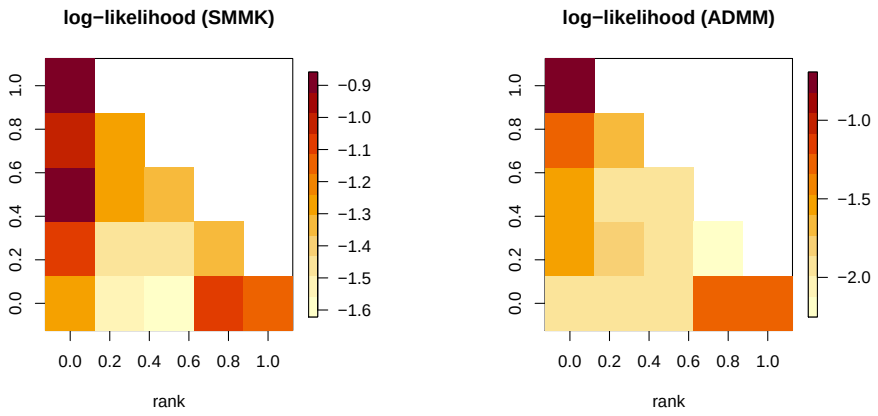
[2] Ulisses M Braga-Neto and Edward R Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

[3] Y. Fan and C. Y. Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 75:531–552, 2013.

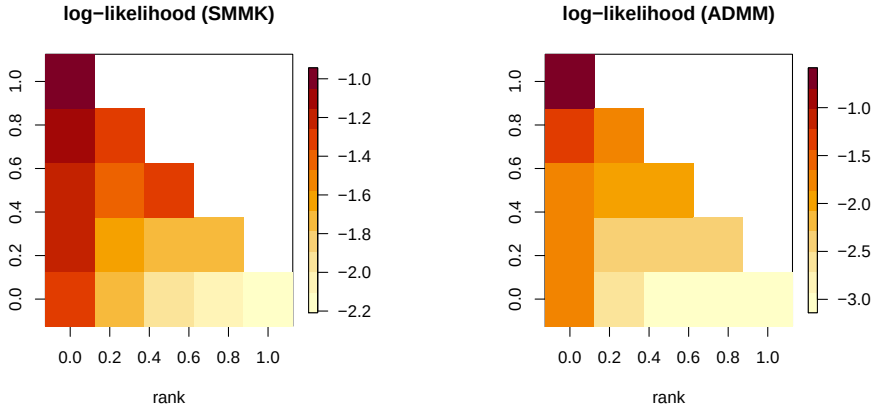
[4] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

[5] Gaël Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, 180:68–77, 2018.

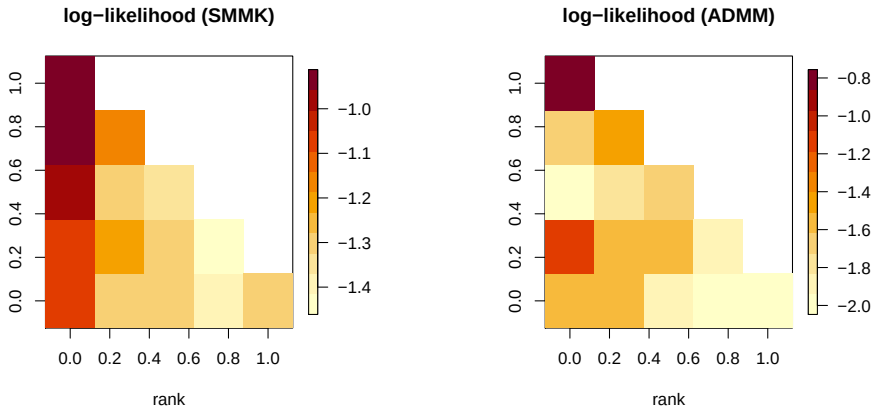
[6] H. Wang, Bo Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 2009.



(a) 2-folded cross validation result.



(b) 5-folded cross validation result.



(c) 10-folded cross validation (LOOCV) result.

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

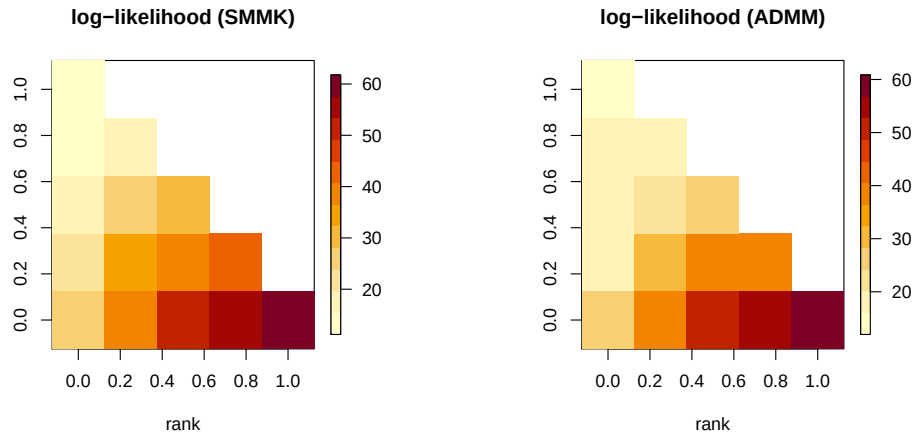


Figure 3: BIC values across different combinations of rank and sparsity according to different algorithms.

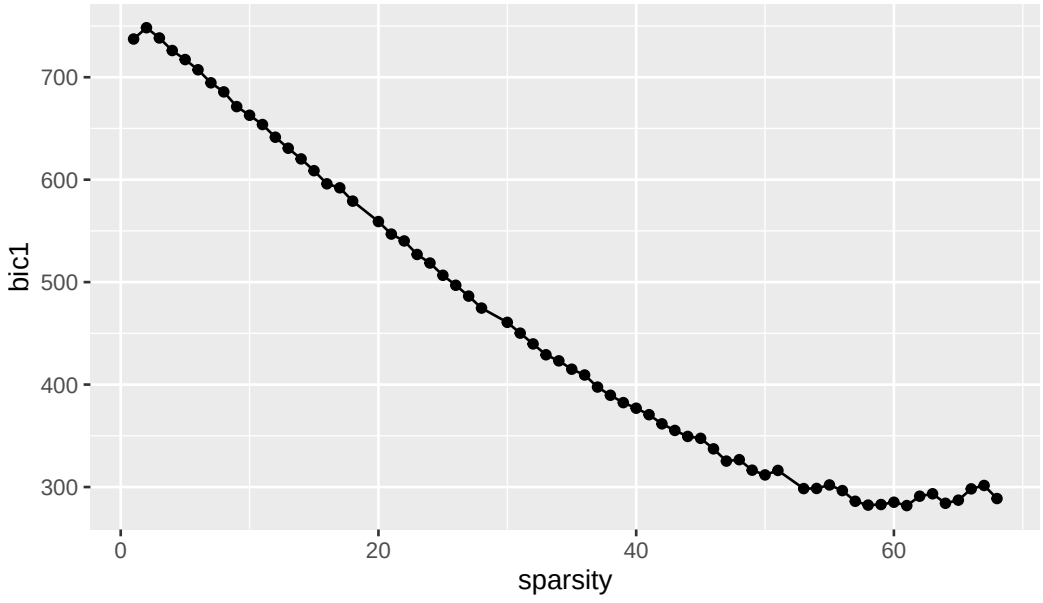


Figure 4: BIC result of VSPLOT brain dataset with fixed rank = 1 and different sparsities.

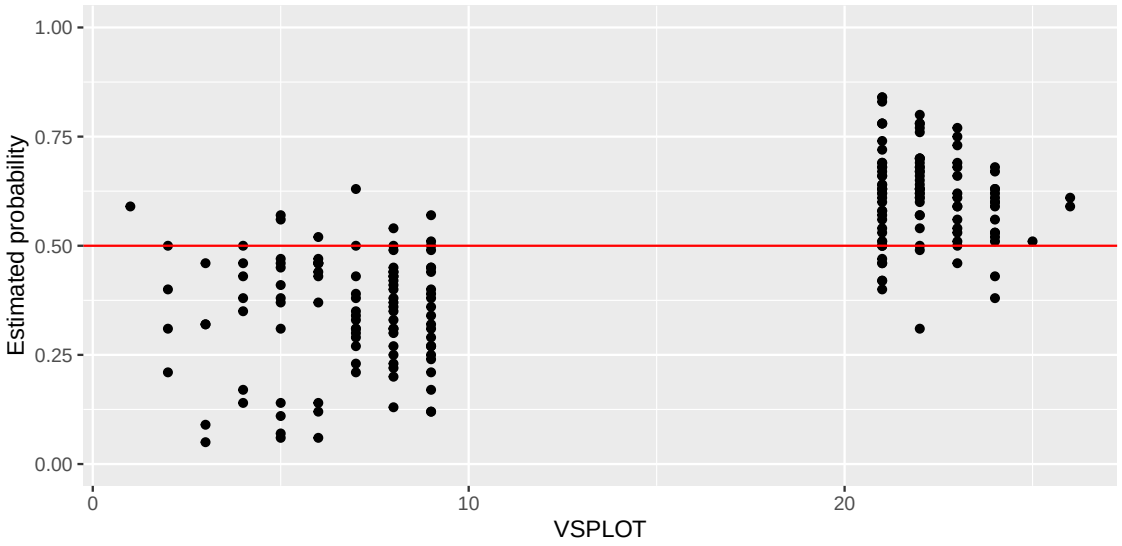


Figure 5: X-axis is VSPLOT scores of individual while y-axis is an estimated probability $\mathbb{P}(Y = 1|\boldsymbol{X})$ of individual. Red horizontal line is 0.5 probability.