# Lasso k-folds CV and repeat n times classification

## Introduction

There are two important points for classification in Metagenome-Wide Association Study (MWAS), one is variable selection that means extract important bacteria, the other is getting better prediction effect, that usually judged by AUC. In this report, the classifier is more robust by adding more procedures depending on the former lasso method. Similarly, the R packages (glmnet and pROC) were used.

## Data info

171 samples collected from community and Research center for mental disorders, includes 81 healthy controls (HC) and 90 schizophrenes (SZ). In SZ, 49 patients are first-episode, and 41 patients are relapse (file name ../../SZData/state171.txt). 360 mOTUs profile which represents bacteria abundance was calculated by analyzing the sequencing data. As the metagenomics data, mOTUs profile contains numerous zero. (file name ../../SZData/motu.0.05.txt)

## Procedure

As a binary classification, just focus on lasso which can do variable selection, the key points is to determine lambda. The former method is just do one k-fold CV, choosing lambda with minim SSE, but the result usually has great contingency. The other way to classify and extract marker is using N times resampled k-fold CV, and then mean test prediction, which used to ROC analysis. In each k-fold CV, one lambda was determine by 5-fold CV, and k-1 parts samples in k-fold CV was used as training set, the surplus one part samples was used as testing set (1).
Following the upper methods, I did 5-fold CV 10 times and in each 5-fold CV, I did 5-fold CV to choose best lambda by setting specific seed. Though these procedure, maybe neither the number of variable nor the robust variable is appropriate. So, filter models and variables is essential. As for model, nonzero coefficient number in samples at least 5, and as for variable, nonzero coefficient variable is at least 50% of the LASSO models (Tab 1), and according to the mean predicted score, the AUC is 78% (Fig 1, A).
Except the original motu profile, I used standardizing motu profile to do classification. I did log10-trainsform and subsequently standardized features by using z-score standardization. The AUC is better than the original result which AUC is 83% (Fig 1, B). And the variables were filtered by the same methods (Tab 2)

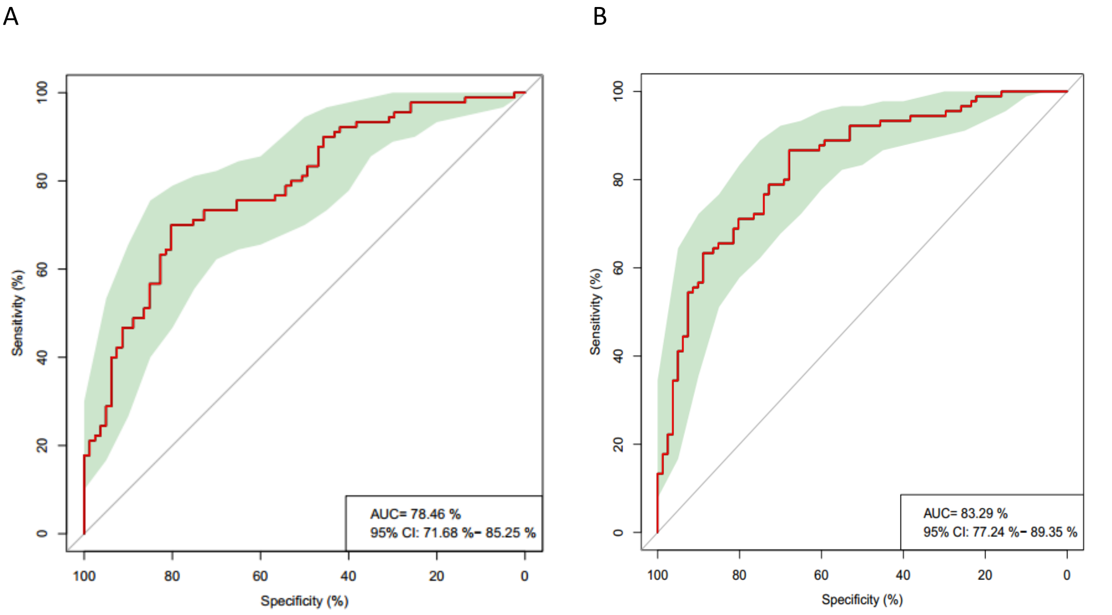
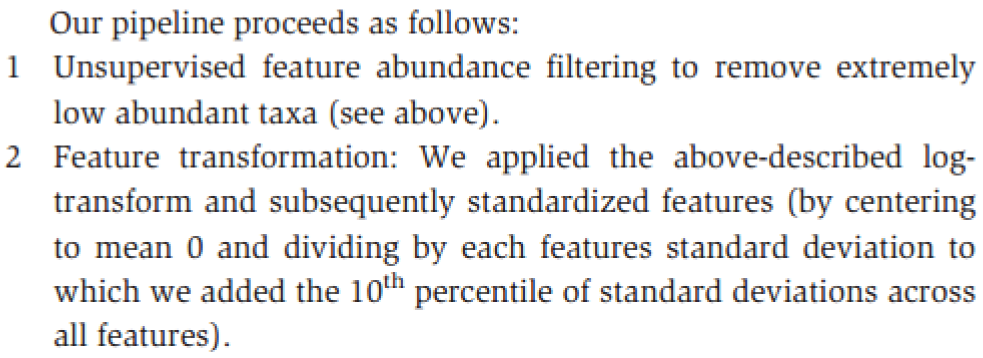


**Fig 1 ROC curve of motu profile.**

**A** ROC curve depending on the original motu profile. **B** ROC curve depending on the standardizing motu profile.

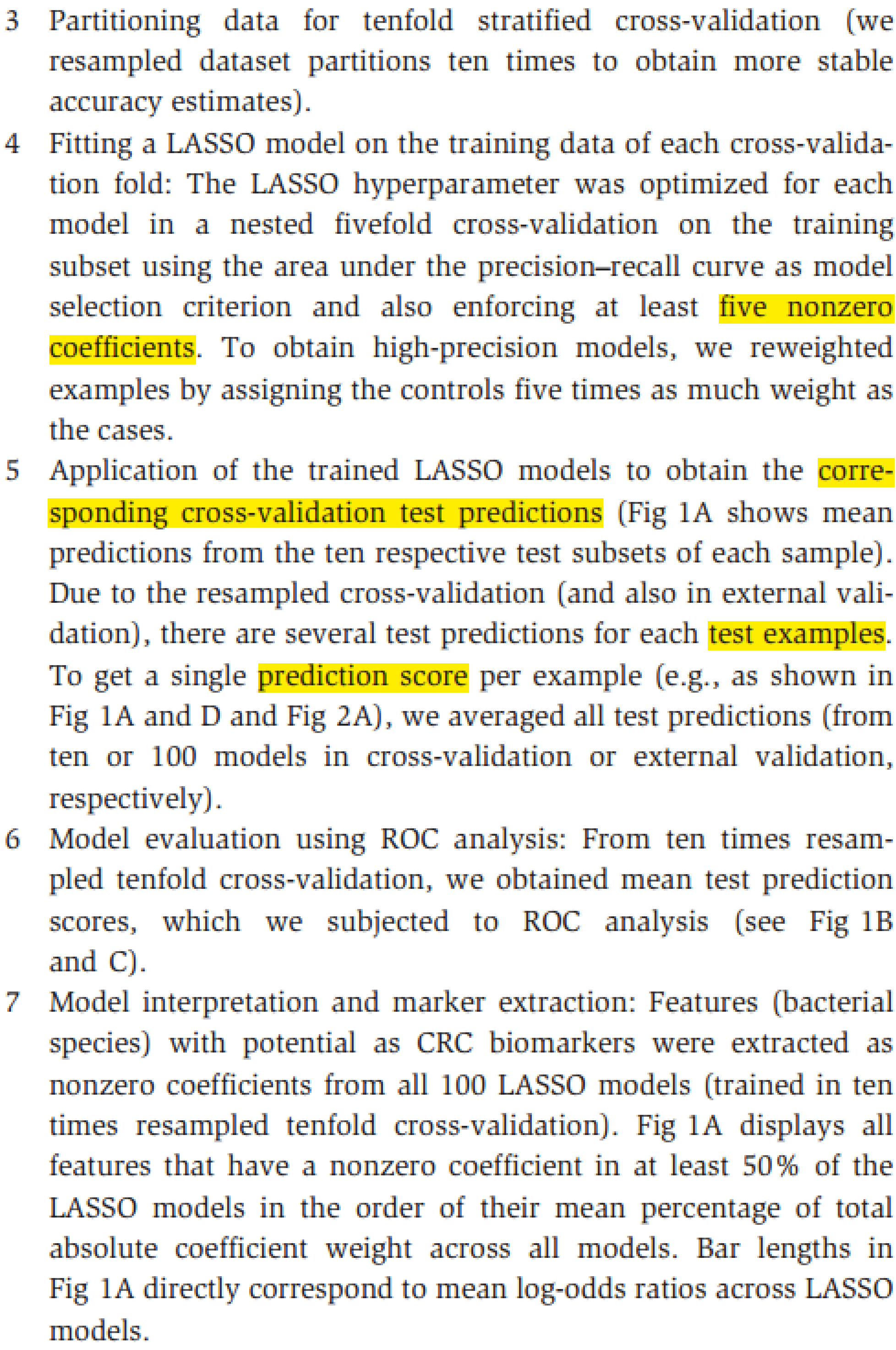
To show the lasso method more clearly, I pasted the screenshot of the reference paper. And the R code is named by Lasso_Classification_Rep.R



Our pipeline proceeds as follows:

1 Unsupervised feature abundance filtering to remove extremely low abundant taxa (see above).

2 Feature transformation: We applied the above-described log-transform and subsequently standardized features (by centering to mean 0 and dividing by each features standard deviation to which we added the 10th percentile of standard deviations across all features).



3 Partitioning data for tenfold stratified cross-validation (we resampled dataset partitions ten times to obtain more stable accuracy estimates).

4 Fitting a LASSO model on the training data of each cross-validation fold: The LASSO hyperparameter was optimized for each model in a nested fivefold cross-validation on the training subset using the area under the precision–recall curve as model selection criterion and also enforcing at least five nonzero coefficients. To obtain high-precision models, we reweighted examples by assigning the controls five times as much weight as the cases.

5 Application of the trained LASSO models to obtain the corresponding cross-validation test predictions (Fig 1A shows mean predictions from the ten respective test subsets of each sample). Due to the resampled cross-validation (and also in external validation), there are several test predictions for each test examples. To get a single prediction score per example (e.g., as shown in Fig 1A and D and Fig 2A), we averaged all test predictions (from ten or 100 models in cross-validation or external validation, respectively).

6 Model evaluation using ROC analysis: From ten times resampled tenfold cross-validation, we obtained mean test prediction scores, which we subjected to ROC analysis (see Fig 1B and C).

7 Model interpretation and marker extraction: Features (bacterial species) with potential as CRC biomarkers were extracted as nonzero coefficients from all 100 LASSO models (trained in ten times resampled tenfold cross-validation). Fig 1A displays all features that have a nonzero coefficient in at least 50% of the LASSO models in the order of their mean percentage of total absolute coefficient weight across all models. Bar lengths in Fig 1A directly correspond to mean log-odds ratios across LASSO models.

# Reference

1. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology. 2014;10:766.