Simulation Result-Real Dataset

Seung Bong Jung

August 2021

In this note, we provide the simulation result for real dataset. We consider breast cancer winsconsin diagnostic dataset. The dataset is available from UCI machine learning data repository. We assess the performance of our proposed estimator, graphical lasso (GL), and sample covariance matrix (Samp) for linear discriminant analysis (LDA). The dataset consist of 212 cases and 357 controls with 30 features, which describe the cell nuclei present in the image of a fine neelde aspirate of breast mass.

We split the dataset into train set with 72 cases and 119 controls and test set with 140 cases and 238 controls. Consider the case as class 1 and the control as class 0. The LDA rule for observation X_i $(i = 1, 2, \ldots, 378)$ in test set is given by

$$\delta(X_i) = \arg\max_{j=0,1} \left\{ X_i^t \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_j - \frac{1}{2} \hat{\mu}_j^t \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_j + \log \hat{\pi}_j \right\},$$

where $\hat{\mathbf{\Sigma}}$ is the estimated covariance matrix based on train set, $\hat{\mu}_j$ is the sample mean of class j among train set, and $\hat{\pi}_j$ is the proportion of class j among train set. We calculate the error rate for test set and replicate this procedure for 10 times. Below the table shows the mean over 10 replications and the value in parentheses denotes the standard deviation.

Proposed (MPP)	Proposed (MAP)	GL	Samp
0.066	**0.066**	0.072	0.077
(**0.016**)	(**0.016**)	(0.017)	(0.019)

TABLE 1: Classification error rate for breast cancer dataset

One may see that our proposed estimators perform better than other estimators.