Low Rank and Generalized Sparse Multi-Task Learning : Data Analysis Example

Youngjin Cho

2021 12 28

Model Description

The model we fit in this example is Low Rank and Generalized Sparse Multi-Task Learning (LGSMTL). For more detail, check LGSMTL2 from the simulation and CCLE data analysis section in the main paper.

Preparation for Fitting

Before fitting the model with example data, please load "Example_Data_Z_neq_I.Rdata", which contains data for this example and run functions in "Example_Code_Z_neq_I.R", which contains functions for fitting LGSMTL.

Data Description

In this example, we use Cancer Cell Line Encyclopedia (CCLE) data for fitting LGSMTL. The source for the CCLE data is https://depmap.org/portal/. To describe the CCLE data, CCLE data consists of 482 cancer cell lines and for each cell line, there are drug resistance responses for 24 drugs and 18988 gene informations. More description for the CCLE data can be found in the main paper. We want to fit resistance responses by gene information in the CCLE data.

Before analyzing the CCLE data, we first do screening for gene information since there are so many genes in the data. See CCLE data analysis section in the main paper for more details on the screening step. After screening the genes, 459 out of 18988 genes remain in the input data. We use them to fit drug resistance responses by LGSMTL.

```r
# Dimension for Gene Information
dim(input_data)
```

```
## [1] 482 459
```

Also, among drug resistance responses for 24 drugs, we screen one drug resistance response. See CCLE data analysis section in the main paper for more details on the screening step. So we use remaining 23 drugs for LGSMTL.

```r
# Dimension for Drug Resistance Response
dim(output_data)
```

```
## [1] 482  23
```

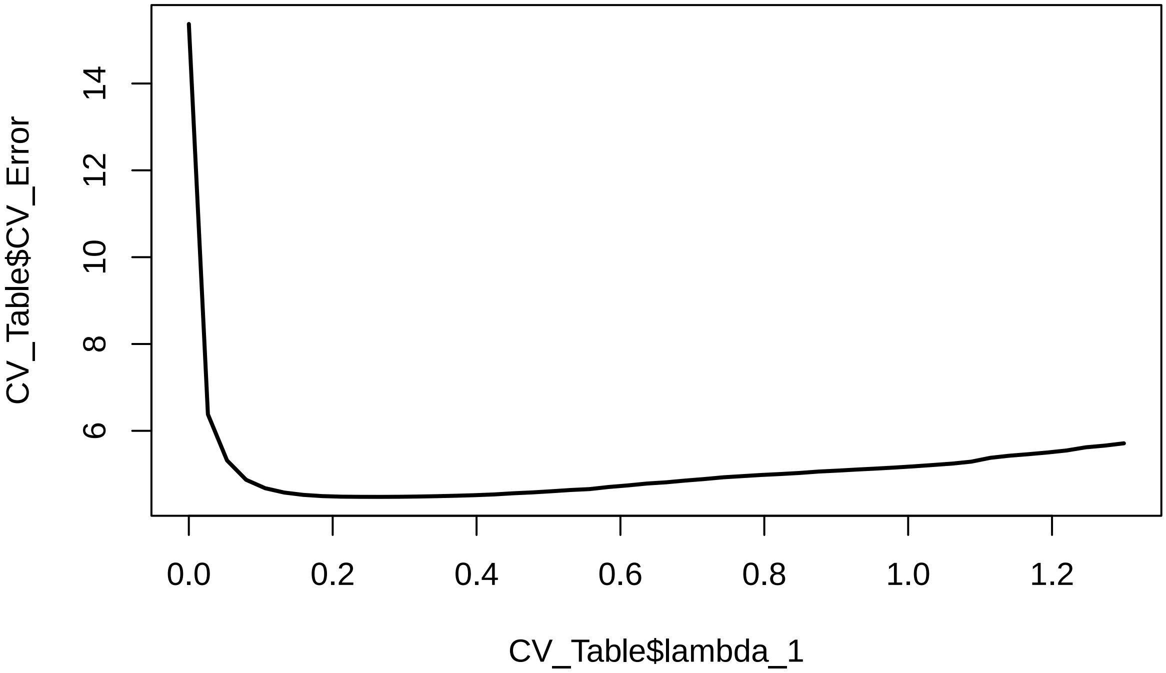
Fitting LGSMTL for the CCLE Data

We fit 482 × 23 drug resistance responses by 482 × 459 gene information. So the coefficient matrix is 459 × 23 dimension matrix. We assume this coefficient matrix is low rank and this low rank structure is due to the correlation among response variables, so we use LGSMTL to fit the coefficient. For tuning parameter selection, the function uses 5-fold Cross validation. The following is fitting function for the LGSMTL.

```
CVfit=CVfit_LGSMTR(input_data,output_data,group,maxgrid=1.3)
```

"input_data" is independent variable for the model and "output_data" is dependent variable for the model. "maxgrid" is maximum value of tuning parameter for doing grid search in Cross Validation. "group" is predifined Cross validation group. As we use 5-fold Cross Validation, we have 5 categories in the group. After fitting model by cross validation, one can check Cross Validation Error like following :

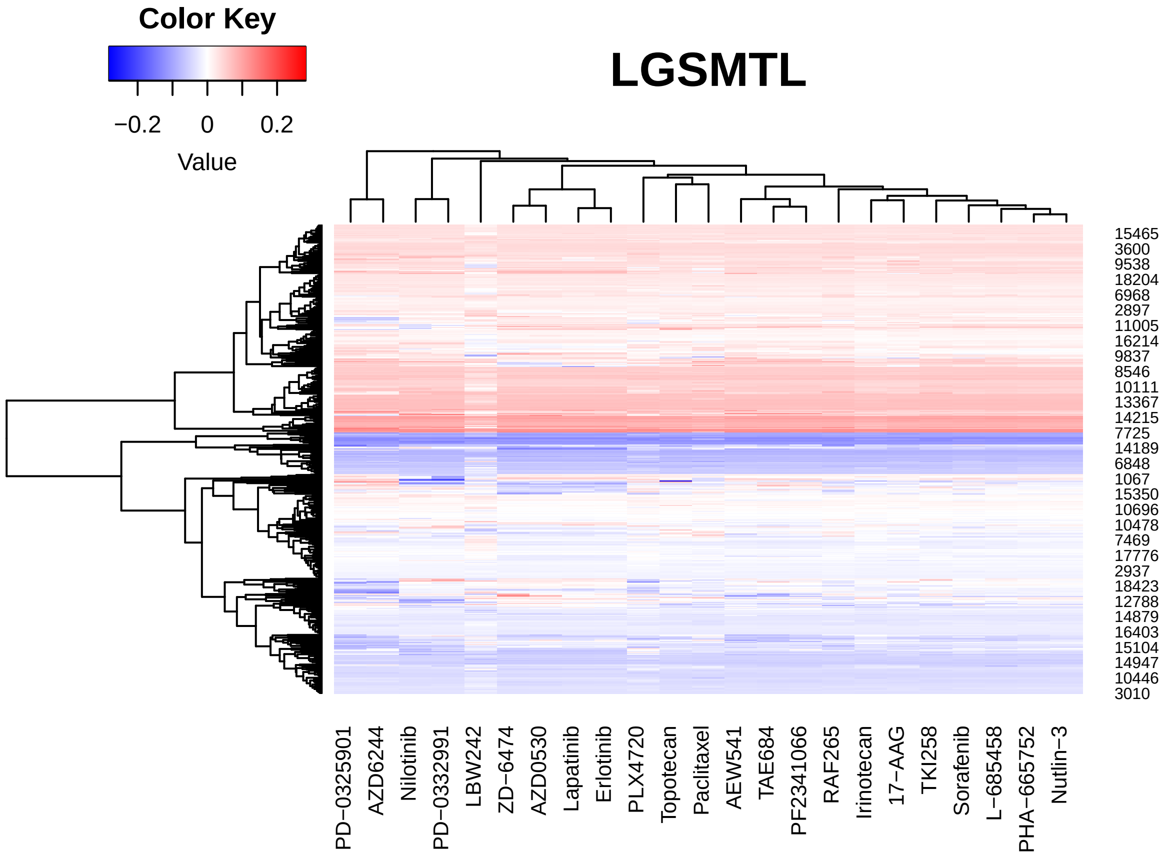
```
CV_Table <- CVfit$‘CV Error‘
plot(CV_Table$lambda_1,CV_Table$CV_Error, type="l", lwd=2)
```



One can see that Cross Validation Error plot shows convex shape, which means the tuning parameters are selected properly.

Now we can check our estimated coefficient matrix. One can use following heatmap function to see the structure of the estimated coefficient.

```
heatmap.2(CVfit$‘Estimated Coefficient‘, col=bluered(1000),
          trace="none", density.info="none", scale="none",
          main="LGSMTL",cex.main=1.5,cexCol=0.8)
```



In the heatmap for coefficient, the columns are for drug resistance responses and the rows are for gene information. One can see that the columns for the estimated coefficient matrix are similar, which means the estimated matrix is low rank.