Network Elastic Net for Identifying Smoking specific gene expression for lung cancer

Contact Information:
Department of Applied Maths and Statistics
Stony Brook University
Stony Brook, NY 11790
barnwal.avinash@stonybrook.edu

Stony Brook University

Avinash Barnwal
Stony Brook University

Abstract

Survival month for non-small lung cancer patients depend upon which stage of lung cancer is present. Our aim is to identify smoking specific gene expression biomarkers in prognosis of lung cancer patients. In this paper, we introduce the network elastic net, a generalization of network lasso that allows for simultaneous clustering and regression on graphs. In network elastic net, we consider similar patients based on smoking cigarettes per year to form the network. We then further find the suitable cluster among patients based on coefficients of genes having different survival month structures and showed the efficacy of the clusters using stage enrichment. This can be used to identify the stage of cancer using gene expression and smoking behavior of patients without doing any tests.

Motivation

- Smokers behaving in similar manner.
- Correlated data forming networks to build separate local models.
- Finding the stage of cancer based on gene expression biomarkers.

Data

TCGA: LUAD, LUSC

- Molecular data from The Cancer Genome Atlas for Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) from cBio Cancer Genomics Portal (http://www.cbioportal.org/), Broad Firehose website (https://gdac.broadinstitute) and from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/).
- Clinical Data - 1024 patients, 77 Features.
  - first nested item
  - second nested item
- Gene Expression Data - 1016 Patients, 19223 Features(Genes).
- Mapped LUSC data - 501 Patients, 19300 Features.
- Mapped LUAD data - 515 Patients, 19300 Features.

Network Lasso

[1] focuses on optimization problems posed on graphs. Consider the following problem on a graph G = (V,E), where V is the vertex set and E is the set of edges:

$$minimize \sum_{i\epsilon v} f_i(x_i) + \sum_{(j,k)\epsilon e} g_{jk}(x_j, x_k) \tag{1}$$

The variables are $x_1, ..., x_n \epsilon R^p$, where $n = |V|$.(The total number of scalar variables is np.) Here $x_i \epsilon R^p$ is the variable at node i, $f_i : R^p \rightarrow R \cup \infty$ is the cost function at node i, and $g_{jk} : R^p \times R^p \rightarrow R \cup \infty$ is the cost function associated with edge(j, k).

$$min \sum_{i\epsilon v} f_i(x_i) + \lambda_1 \sum_{(j,k)\epsilon e} w_{jk}||x_j - x_k||_2 \tag{2}$$

Our proposed regularizer can be solved by a general alternating direction method of multipliers (ADMM) based solver.It consists of the following steps (for each iteration k).

$$\begin{aligned} x^{k+1} &= \arg\min_x L_p(x, z^k, u^k) \\ z^{k+1} &= \arg\min_z L_p(x^{k+1}, z, u^k) \\ u^{k+1} &= u^k + (x^{k+1} - z^{k+1}) \end{aligned} \tag{3}$$

Network Elastic-Net

Modeling: Network Lasso, Isotropic Total Variation, Anisotropic Total Variation

- [3] focuses on using combination anisotropic and istropic total variations regularization.
- Isotropic Total Variation is equivalent to group lasso [5], also known as $l_{1,2}$ regularization.
- Anisotropic Total Variation is equivalent to fused lasso [4], also known as l_1 regularization.

Here regularization is the combination of l_1 and $l_{1,2}$.

$$\begin{aligned} min \sum_{i\epsilon v} f_i(x_i) + \lambda_1 \sum_{(j,k)\epsilon e} w_{jk}||x_j - x_k||_2 + \\ \lambda_2 \sum_{(j,k)\epsilon e} w_{jk}||x_j - x_k|| \end{aligned} \tag{4}$$

ADMM

- Same updates except z-update.
- Coordinate Descent for z-update.

Simulation

We have created high dimensional synthetic data having clustered and orthogonal coefficients. First, we have generated the predictors such that $x_{ij} \sim Unif(-1, 1)$, j = 1,...,10 and i = 1,...,100 and $e_i \sim N(0, 1)$ and corresponding response variable have been generated using below model :-

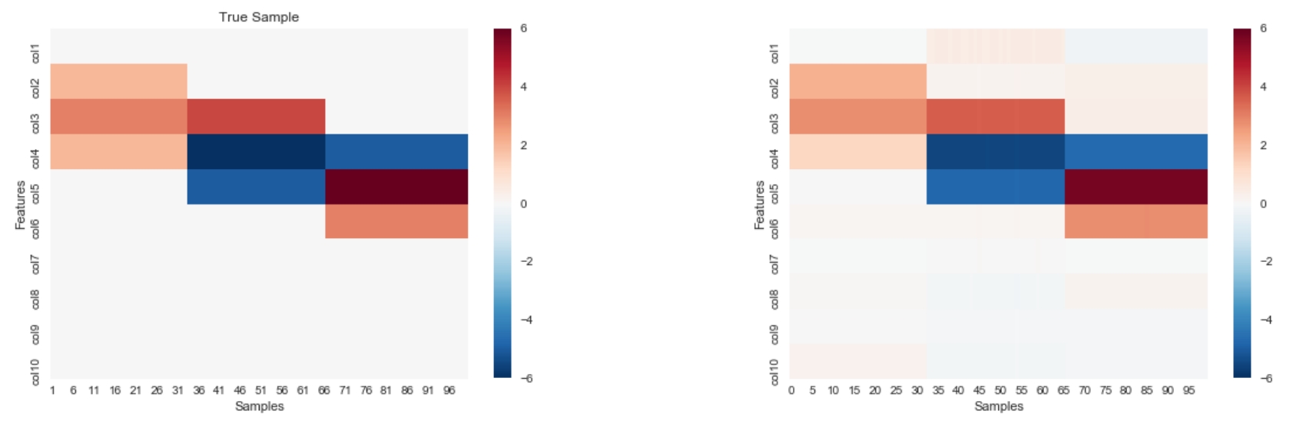
$$y_i = 2 * x_{i2} + 3 * x_{i3} + 2 * x_{i4} + 0.1 * e_i, i = 1, .., 33 \tag{5}$$

$$y_i = 4 * x_{i3} - 6 * x_{i4} - 5 * x_{i5} + 0.1 * e_i, i = 34, ..., 66 \tag{6}$$

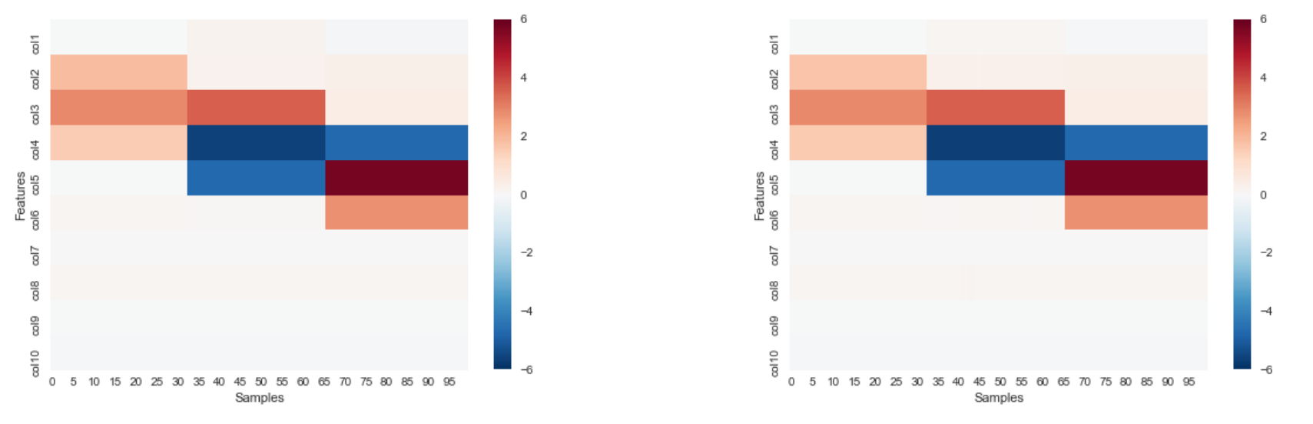
$$y_i = -5 * x_{i4} + 6 * x_{i5} + 3 * x_{i6} + 0.1 * e_i, i = 67, ..., 100 \tag{7}$$

Above 3 equations are 3 networks.

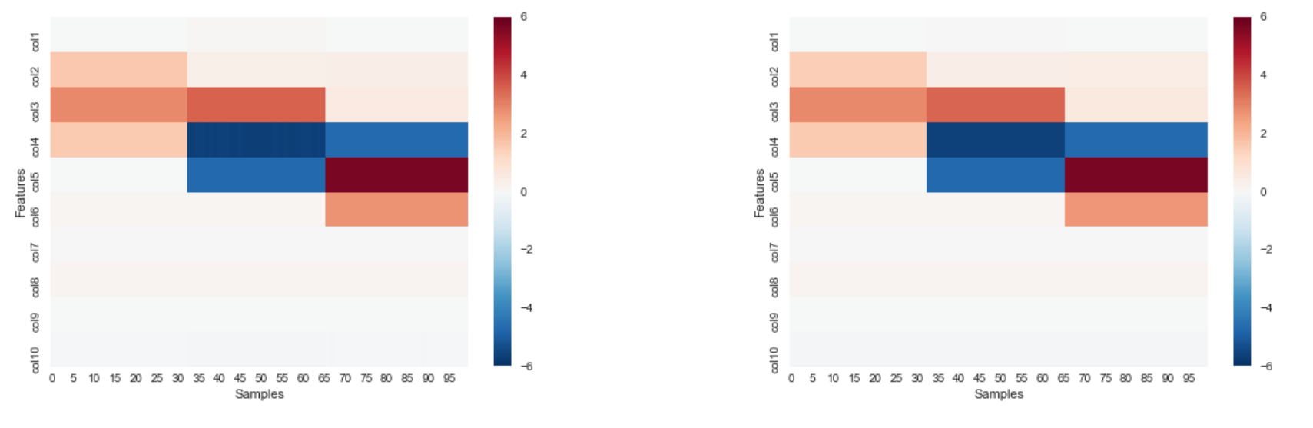
Weight Matrix - 95% density with 1 and 0 for 5%.



(a): True Sample and λ = 1.12,α = 0



(b): λ = 1.12,α = 0.4 and λ = 1.12,α = 0.6



(c): λ = 1.12,α = 0.8 and λ = 1.12,α = 1

Survival Model-Accelerated Failure Time(AFT)

Modeling: Accelerated Failure Time

[2] discusses Accelerated Failure Time being weighted least square regression, where,
$Y_{(1)} \leq ... \leq Y_{(n)}$ - order statistics of $Y_i's$.
$\delta_{(1)}, ..., \delta_{(n)}$ - associated censoring indicators.
$X_{(1)}, ..., X_{(n)}$ - associated covariates. and it is further mean adjusted leading to below equations.
Denote

$$X^*_{(i)} = (nw_i)^{\frac{1}{2}}(X_{(i)} - \overline{X}_w) \tag{8}$$

$$Y^*_{(i)} = (nw_i)^{\frac{1}{2}}(Y_{(i)} - \overline{Y}_w) \tag{9}$$

Therefore weighted loss function is

$$L = \sum_{i=1}^{n} (Y^*_{(i)} - X^*_{(i)}\beta)^2 \tag{10}$$

For Network Elastic Net

$$L = \sum_{i=1}^{n} (Y^*_{(i)} - X^*_{(i)}\beta)^2 + \tag{11}$$

$$\lambda(1 - \alpha) \sum_{(j,k)\epsilon E} w_{jk}||\beta_j - \beta_k||_2 + \lambda\alpha \sum_{(j,k)\epsilon E} w_{jk}||\beta_j - \beta_k||_1 \tag{12}$$

Cross-Validation

Robust Cross-Validation is important step to find the hyper-parameters. We have split the data into 80:20 where Training data is 80% and Testing data is 20%. For performance measure, [2] uses AIC.

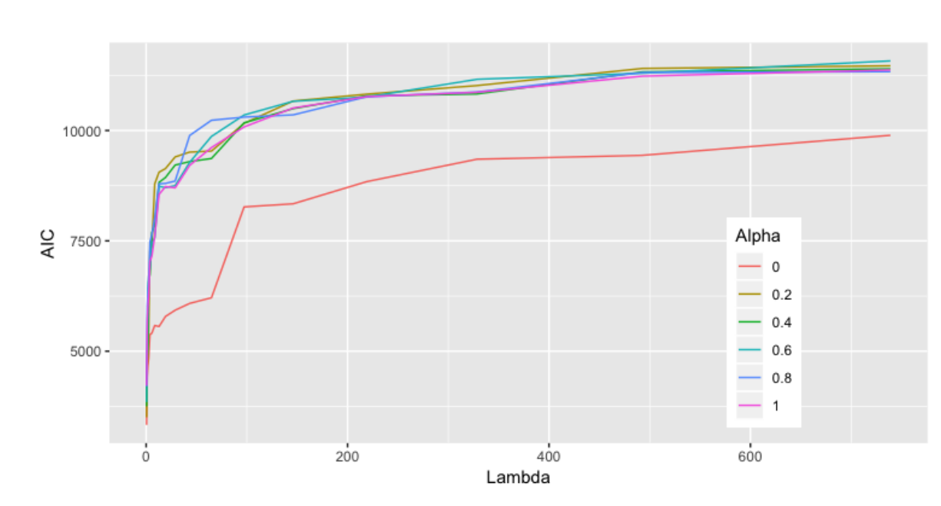
$$AIC - Score = nlog(CV Score) + 2K \tag{13}$$

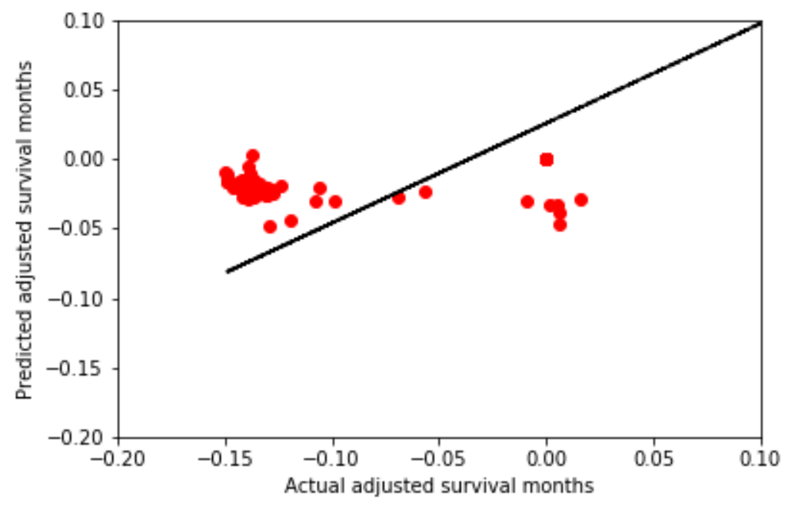
where CV Score is the loss for test data and K is corresponding non-zero coefficients.

Result

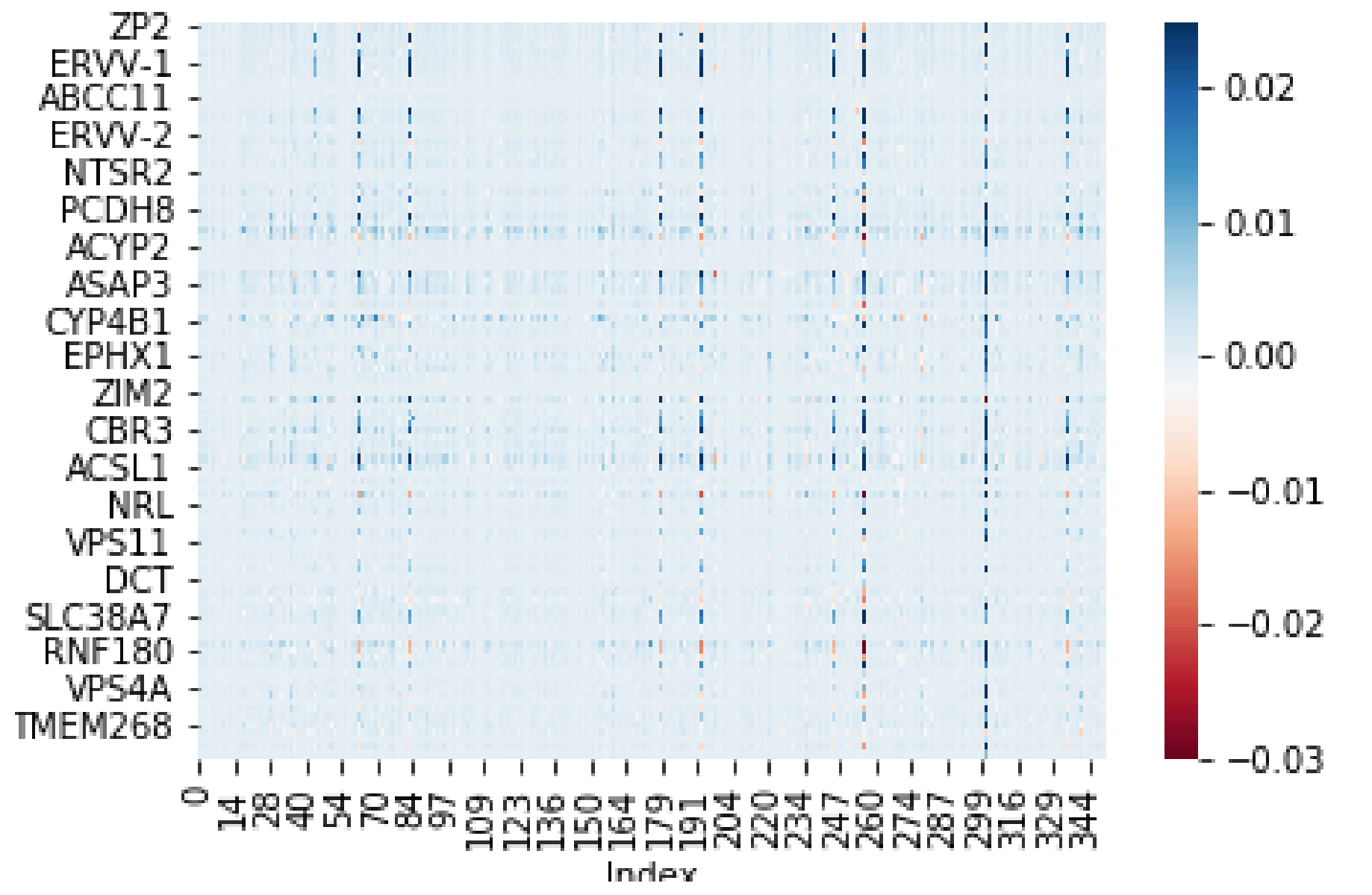
LUAD Cancer Type - Top 100 Genes correlated with Survival Months

- Hyperparameter(λ) performance
- Predicted vs Actual Adjusted Survival Months





- Coefficient matrix.



- Inferring Stage of Cancer based on coefficient of gene variables.
- Clustering based on Coefficient of gene variables.

Count of Stage of Cancer vs Cluster

Cluster	Stage 1	Stage 2	Stage 3	Stage 4
1	53	26	12	4
2	3	2	1	0
3	36	13	13	4
4	45	20	17	6
5	13	6	6	1

Significance of Stage of Cancer vs Cluster

Cluster	Stage 1	Stage 2	Stage 3	Stage 4
1	0.00	0.00	0.86	0.77
2	0.42	0.32	1.00	1.00
3	0.00	0.59	0.11	0.49
4	0.00	0.11	0.04	0.13
5	0.14	0.60	0.24	1.00

Conclusion

- λ = 0.5 and α = 0 shows least **AIC** - 3331.
- Correlation between Actual adjusted survival months and predicted adjusted survival months is 0.77.
- 31 Significant Genes having norms greater than 0.05.
- Cluster 1 belongs to Stage 1 or Stage 2, Cluster 3 belongs to Stage 1 and Cluster 4 belongs to Stage 1 or Stage 3.

References

[1] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. pages 387–396, 2015.

[2] Jian Huang, Shuangge Ma, and Huiliang Xie. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62(3):813–820, 2006.

[3] Yifei Lou, Tieyong Zeng, Stanley Osher, and Jack Xin. A weighted difference of anisotropic and isotropic total variation model for image processing. *SIAM J. Imaging Sciences*, 8(3):1798–1823, 2015.

[4] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.

[5] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.