Identification of smoking specific gene expression biomarkers for lung cancer using elastic network lasso model

immediate

January 31, 2018

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. We develop an algorithm based on the Alternating Direction Method of Multipliers (ADMM) to solve this problem in a distributed and scalable manner, which allows for guaranteed global convergence even on large graphs. In network elastic net, we consider similar patients based on smoking cigarettes per year,forming cluster behave accordingly. We then further find the suitable cluster among patients based on coefficients of genes having different survival month structures.
Keywords: Convex Optimization, Network Theory, Regression, Elastic Net

1. Introduction

One of the key challenge in molecular medicine is to personalize the feature selection for each data point on sample space. This can be treated as local feature selection and prediction problem. Recently, network lasso has been introduced where coefficient difference is penalized with l2 norm given graph structure. It has already been observed that Elastic net works better than lasso with real world data for multivariate regression problems. We have introduced elastic net norm for network regression and tested the results for Survival months for non-small lung cancer patients for early stage cancer patients.

2. Formulation and Methodology

We are focusing 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:

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

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). We use extended (infinite) values off i and g_{jk} to describe constraints on the variables, or pairs of variables across an edge,respectively.Our focus will be on the special case in which the fiare convex, and $g_{jk}(x_j, x_k) = \lambda_1 w_{jk} ||x_j - x_k||_2 + \lambda_2 w_{jk} ||x_j - x_k||$, with $\lambda \geq 0$ and user-defined weights $w_{jk} \geq 0$:

$$minimize \sum_{i \epsilon v} f_i(x_i) + \lambda_1 \sum_{(j,k) \epsilon e} w_{jk} ||x_j - x_k||_2 + \lambda_2 \sum_{(j,k) \epsilon e} w_{jk} ||x_j - x_k|| \quad (2)$$

(2)The edge objectives penalize differences between the variables at adjacent nodes, where the edge between nodes i and j has combination of l_2 and l_1 norm having weights $\lambda_1 w_{ij}$ and $\lambda_2 w_{ij}$. w_{ij} can be considered as the graph property having similar nodes with more penalization, and λ_1 and λ_2 as an overall parameter that scales the edge objectives relative to the node objectives. We call problem (2) the network regression with elastic net problem, since the edge cost is a sum of $l1$ and $l2$ norms of differences of the adjacent edge variables . Hallac et. al has showed that network lasso is a convex optimization problem,and we have added a convex function part to it which maintains property of convex optimization problem. Here , problem requires scalable solution with large nodes and variables. Our proposed regularizer can be solved by a general alternating direction method of multipliers (ADMM) based solver.

ADMM consists the following steps (for each iteration k)

$$x^{k+1} = \arg\min_x L_p(x, z^k, u^k) \quad (3)$$

$$z^{k+1} = \arg\min_z L_p(x^{k+1}, z, u^k) \quad (4)$$

$$u^{k+1} = u^k + (x^{k+1} - z^{k+1}) \quad (5)$$

Each step can be described as x - update, z - update, and u - update.

As there will not be any changes in x-update and u-update compared to network lasso.But z-update is having extra $l1$-norm which is solved using soft-thresholding operator.

We have created the graph structure using the similarity between humans using 19196 genes. We have omitted genes having high number of missing values. This leads to weight of similarity based on correlation between two set of genes for corresponding patients.

Following is the algorithm for ADMM Step

Algorithm 1 ADMM Steps

repeat

$x_i^{k+1} = \underset{x_i}{\operatorname{argmin}}\big(f_i(x_i) + \sum_{j \in N(i)} (\rho/2)||x_i - z_{ij}^k + u_{ij}^k||_2^2\big)$

$a = x_i^{k+1} + u_{ij}^k, b = x_j^{k+1} + u_{ji}^k$

$c_1 = \lambda_1 * (1-\alpha) * w_{ij}, c_2 = \lambda_1 * (\alpha) * w_{ij}$

$\mu_1 = ||\rho * (a-b) - 2 * c_2||_2/\rho - 2 * c_1/\rho$

$\mu_2 = ||\rho * (a-b) + 2 * c_2)/\rho - 2 * c_1/\rho$

$\gamma_1 = \mu_1 * c_2/(2 * c_1 + \mu_1 * \rho)$

$\gamma_2 = \mu_2 * c_2/(2 * c_1 + \mu_2 * \rho)$

$\theta_1 = 0.5 + \mu_1 * \rho/(4 * c_1 + 2 * \mu_1 * \rho), \theta_2 = 0.5 + \mu_2 * \rho/(4 * c_1 + 2 * \mu_2 * \rho)$

until $||r^k||_2 \leq \epsilon^{pri}; ||s^k||_2 \leq \epsilon^{dual}$

Following is the algorithm for Regularization path

Algorithm 2 Regularization Steps

for Weight Type $\in$ {Euclidean,Correlation,Diffusion Map} **do**

  for $\alpha \in$ [0,1] **do**

    initialize Solve for x^*, u^*, z^* at $\lambda = 0$.

    repeat

      set $\lambda := \gamma\lambda; \gamma \geq 1$;

      Use Algorithm 1 to solve for $x^*(\lambda), u^*(\lambda), z^*(\lambda)$.

    until $x^*(\lambda) = x^*(\lambda_{previous})$

    return $x^*(\lambda)$ for λ from 0 to $\lambda_{critical}$

  end for

end for

3. Experiments

In this section , we first illustrate our proposed method on synthetic data and then we perform the survival month prediction and clustering of survival months using coefficients of predictors and then for TCGA datasets ,predicting survival months for LUAD and LUSC lung cancers using gene expression variables.

3.1. Synthetic experiments

We have used 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 :-

$$y_i = 2*x_{i2}+3*x_{i3}+2*x_{i4}+0.1*e_i, i = 1, .., 33 \quad (6)$$

$$y_i = 4*x_{i3}-6*x_{i4}-5*x_{i5}+0.1*e_i, i = 34, ..., 66 \quad (7)$$

$$y_i = -5*x_{i4}+6*x_{i5}+3*x_{i6}+0.1*e_i, i = 67, ..., 100 \quad (8)$$

Lets consider the first equation corresponds to first cluster , second equation corresponds to second cluster and third equation corresponds to third cluster. We have one more requirement here, link function , its very important to create the link function such that first cluster has high density among themselves but low density compared to other cluster , same holds for other cluster.

We have generated the link function using following steps:-

1. Create Sparse matrix R $\in \{0, 1\}^{100x100}$ with density 1%.
2. Create Dense matrix R $\in \{0, 1\}^{33x33}$ with density 95%.
3. Create Dense matrix R $\in \{0, 1\}^{34x34}$ with density 95%.
4. Replace corresponding dense matrix in the sparse matrix such that first cluster has high density among themselves but low density compared to other clusters and similar properties for other two clusters.

We experimentally set the regularization parameter for the proposed method to $\alpha \in$ {0,0.2,0.4,0.6,0.8,1} and $\lambda = 1.12$. For the network Lasso, $\alpha = 0$. Moreover, in we can also regularize the coefficients as well. First part of charts show estimated coefficients using our proposed method without coefficients regularization and second part of charts show estimated coefficients using our proposed method with coefficients regularization. We have used $\mu = 0.1$ for coefficient regularization.

Clearly , we can see that estimated coefficients have been recovered for each α and for regularized coefficients $\alpha = 1$ works best.

3.2. Predictions of Survival Months for LUAD and LUSC Cancer

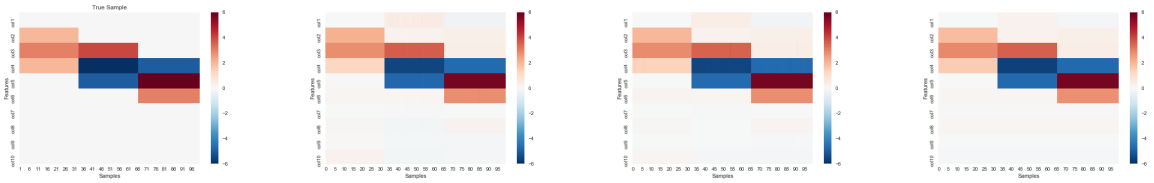
We perform the survival month prediction and clustering of survival months using gene expression data.We have used TCGA dataset having lung cancer patients with clinical variables and corresponding gene expression data. There are 325 patients in LUAD lung cancer and 423 patients in LUSC lung cancer. We have 19196 genes.

Firstly, we have normalized the numeric variables Survival months.We then divided the data into 80:20 to train the model. For Survival months , we have multiplied normalized survival months with 10. We have weight measures based on number of smoking cigarettes per year.If w_i is number of smoking cigarettes per year for $patient_i$ and similarly w_j corresponding for $patient_j$.

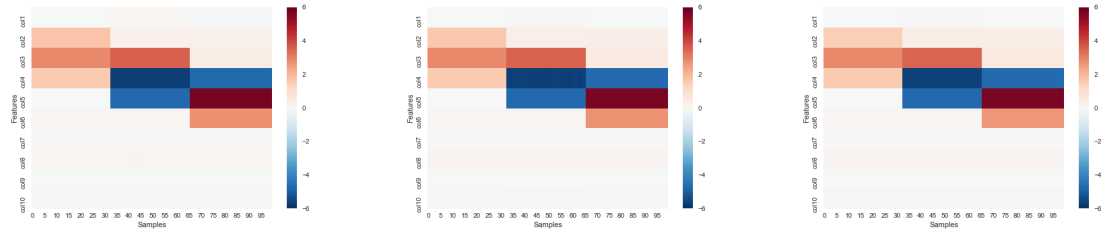
for weight measure

$$w_{ij} = \begin{cases} \frac{1}{w_i - w_j}, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

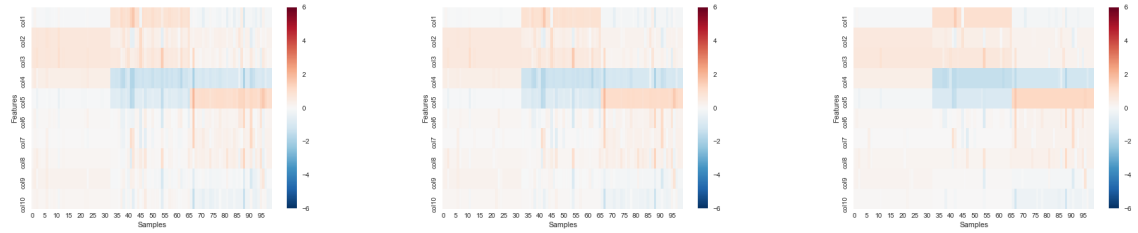
We have kept only 100 top correlated genes



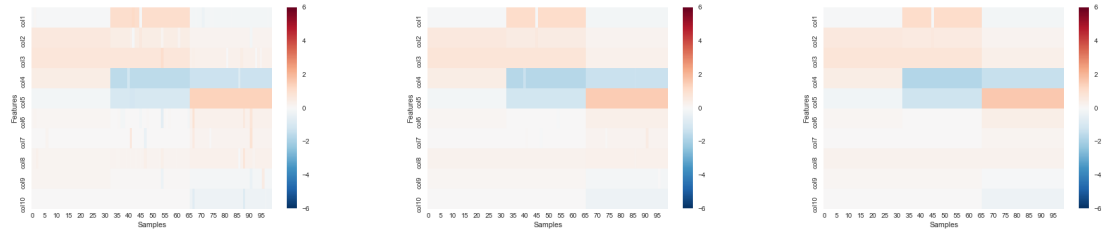
(a) True pattern. (b) λ =1.12 , $\alpha = 0$ and $\mu = 0$. (c) λ =1.12 , $\alpha = 0.2$ and $\mu = 0$. (d) λ =1.12 , $\alpha = 0.4$ and $\mu = 0$.



(e) λ =1.12 , $\alpha = 0.6$ and $\mu = 0$. (f) λ =1.12 , $\alpha = 0.8$ and $\mu = 0$. (g) λ =1.12 , $\alpha = 1$ and $\mu = 0$.

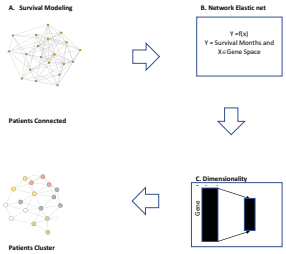


(a) λ =1.12 , $\alpha = 0$ and $\mu = 0.1$ (b) λ =1.12 , $\alpha = 0.2$ and $\mu = 0.1$ (c) λ =1.12 , $\alpha = 0.4$ and $\mu = 0.1$



(d) λ =1.12 , $\alpha = 0.6$ and $\mu = 0.1$ (e) λ =1.12 , $\alpha = 0.8$ and $\mu = 0.1$ (f) λ =1.12 , $\alpha = 1$ and $\mu = 0.1$

Figure 2: Simulation of Network Elastic net without coefficients Regularization



(a) WorkFlow

with survival months among 19196 as predictors for each cancer types and lets call it $gene_1$, $gene_2$, ..., $gene_{100}$.

Optimization Parameter and Objective Function For each patient ,we solve for $x_i = [a_{i1}\ a_{i2}...\ a_{i100}]^T$,which gives us the coefficients of the regressors. The survival months estimate is given by

$$y_i = a_{i1} \cdot gene_1 + a_{i2} \cdot gene_2 + ... + a_{i3} \cdot gene_1 00 + c_i$$

,where the constant offset c_i is the "baseline". The objective function for each patient then becomes $f_i = ||Survival\ Month_i - y_i||_2^2$ where , $SurvivalMonth_i$ is the actual survival month for $patient_i$.
To predict the survival month on the test set, we connect each new patient to the 5 nearest patient based on weight measures defined above. We then infer the value of x_j at for $patient_j$ by solving problem , and we use this value to estimate the new survival month for $patient_j$.

We solve for x_j

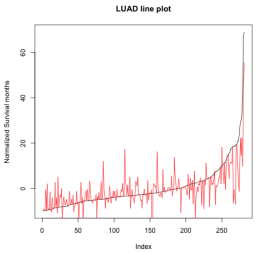
$$min \sum_{k \epsilon N(j)} w_{jk} ||x_j - x_k||_2 \qquad (10)$$

The inference for new patient is being used by Hallac et al.,2015. which is a weber problem.

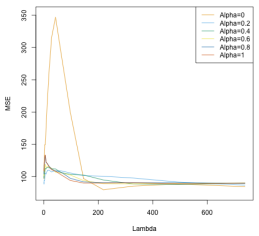
We have taken MSE(mean square error) as a performance metric to decide which model is best. For each α in Network Elastic net we have shown the regulation path and for λ having least MSE in test data, fitted model performance is shown for training dataset.

Following is the flow chart of the method involved:-

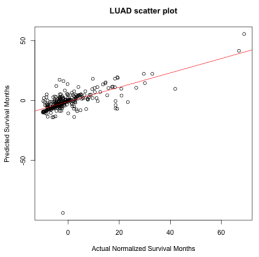
Results for LUAD and LUSC Cancer type:-



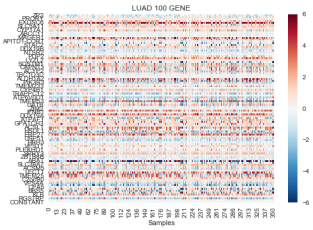
(a) LUAD Cancer- Actual vs Predicted normalized survival months



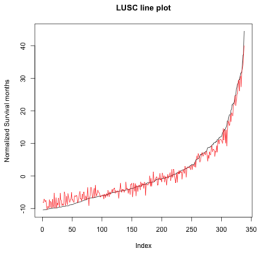
(c) LUAD Cancer - Regularization Path



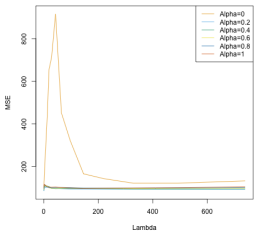
(b) LUAD Cancer - Scatter plot between actual and predicted survival months



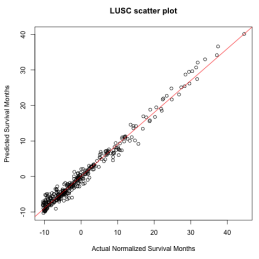
(d) LUAD Cancer - Coefficients Heat Map



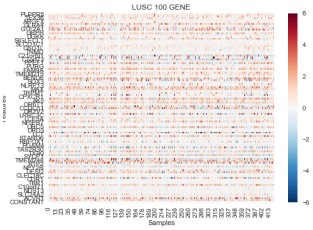
(e) LUSC Cancer- Actual vs Predicted normalized survival months



(g) LUSC Cancer - Regularization Path



(f) LUSC Cancer - Scatter plot between actual and predicted survival months



(h) LUSC Cancer - Coefficients Heat Map

For LUAD Cancer - $\alpha = 0$ and $\lambda = 218.5$ has lead to least MSE $= 79.62$ and for LUSC Cancer - $\alpha = 0.2$ and $\lambda = 0.75$ has lead to least MSE $= 84.54$.

3.3. Clustering of Patients - LUAD and LUSC Cancer Next, we have clustered the patients in training dataset based on gene expression coefficient calculated from Network Elastic Net. Characteristics of clusters can further be validated based on external variables such as Cancer Stage,smoking stages and Km-plot. We have used smoking stage to identify the cluster

characters. Smoking stage is ranged from 2 to 5.

LUAD - Cancer Stage					
Cluster	1	2	3	4	Total
1	21	4	4	2	31
2	40	23	21	6	90
3	15	9	6	2	32
4	36	10	9	3	58
5	15	8	4	1	28
6	23	13	3	3	42
Total	150	67	47	17	281

LUAD - Smoking Stage					
Cluster	2	3	4	5	Total
1	12	6	13	0	31
2	36	16	38	0	90
3	2	21	9	0	32
4	9	26	23	0	58
5	2	13	12	1	28
6	12	4	25	1	42
Total	73	86	120	2	281

LUSC - Cancer Stage					
Cluster	1	2	3	4	Total
1	6	5	2	0	13
2	45	35	21	4	105
3	22	14	9	0	45
4	54	39	12	0	105
5	27	14	10	1	52
6	9	6	3	0	18
Total	163	113	57	5	338

LUSC - Smoking Stage					
Cluster	2	3	4	5	Total
1	2	3	8	0	13
2	27	13	63	1	104
3	19	3	23	0	45
4	29	21	54	0	104
5	16	10	26	0	52
6	4	8	5	1	18
Total	97	58	179	2	336

To identify which cluster corresponds to which stage , we have used similar to gene enrichment strategy where for each cluster and each stage we have calculated corresponding P-value and if P-Value is smaller than <0.05 then cluster represents that stage. Following are the Stage Enrichment for LUAD and LUSC cancer for Cancer and Smoking Stages.

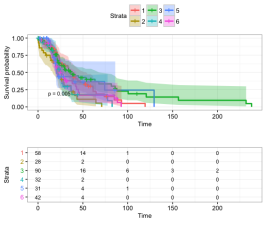
LUAD-Cancer Stage				
Cluster	1	2	3	4
1	0.00	0.63	1.00	0.67
2	0.02	0.02	0.00	0.26
3	0.18	0.15	0.41	0.68
4	0.00	1.00	0.52	1.00
5	0.04	0.20	0.78	1.00
6	0.01	0.03	0.33	0.46

LUAD - Smoking Stage				
Cluster	2	3	4	5
1	0.74	0.41	0.07	1.00
2	0.10	1.00	0.00	0.42
3	0.00	0.24	0.02	1.00
4	0.04	0.01	0.00	1.00
5	0.10	0.20	0.02	1.00
6	1.00	0.00	0.62	0.10

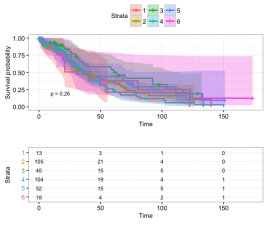
LUSC-Cancer Stage				
Cluster	1	2	3	4
1	0.37	0.33	1.00	1.00
2	0.01	0.01	0.01	0.01
3	0.02	0.27	0.17	1.00
4	0.00	0.00	1.00	0.60
5	0.00	0.61	0.19	0.51
6	0.13	0.41	0.73	1.00

LUSC - Smoking Stage				
Cluster	2	3	4	5
1	0.00	0.33	0.02	1.00
2	0.44	0.45	0.00	1.00
3	0.08	0.55	0.05	1.00
4	0.04	0.50	0.00	1.00
5	0.33	0.17	0.03	0.22
6	1.00	0.00	0.62	0.10

Next , we have also plotted the KM Plot for each cluster to understand the cluster characteristics. Following are the KM Plots for LUSC and LUAD lung cancer clusters:-

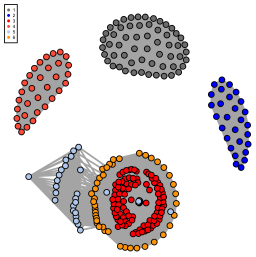


(a) LUAD Cancer- KM Plot

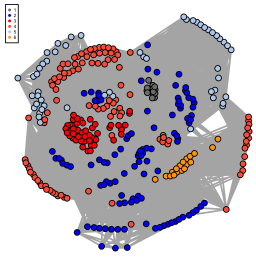


(b) LUSC Cancer - KM Plot

It is also very useful to visualize the network of patients based on gene coefficients. Here, patients are nodes and edge weight is defined based on correlation between gene coefficients.We have calculated weight of the edge based on the correlation of gene coefficients.



(a) LUAD Cancer- Network Cluster



(b) LUSC Cancer - Network Cluster

4. Results

References

[1] David Hallac, Jure Leskovec, Stephen Boyd *Network Lasso: Clustering and Optimization in Large-Graphs*. In KDD,2015.

[2] Makoto Yamada, Koh Takeuchi, Tomoharu Iwata, John Shawe-Taylor, Samuel Kaski *Localized Lasso for High-Dimensional Regression*. arXiv:1603.06743, 2016

[3] Robert. Tibshirani*Regression shrinkage and selection via the Lasso*. Journal of the Royal Statistical Society,Series B, 58(1):267–288, 1996

[4] Hui Zou, Trevor Hastie *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society,Series B, 67(2):301–320, 2005.

[5] Rie Kubota Ando and Tong Zhang*A frame work for learning predictive structures from multiple tasks and unlabeled data*. Journal of Machine Learning Research,6(Nov):1817–1853,2005

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J.Eckstein *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Foundations and Trends in Machine Learning, 3:1–122, 2011

[7] S. Boyd and L. Vandenberghe *Convex Optimization* .Cambridge University Press, 2004

[8] J. Leskovec and R. Sosic. *Snap.py: SNAP for Python* .http://snap.stanford.edu, 2014

Appendix

Analytical solution to Z-Update

Following is the equation having z-variables

$$minimize\ \lambda * (1-\alpha) * w_{ij}||z_{ij} - z_{ji}||_2 + \lambda * (\alpha) * w_{ij}||z_{ij} - z_{ji}||_1 + (\frac{\rho}{2})(||x_i^{k+1} - z_{ij} + u_{ij}^k||_2^2 + ||x_j^{k+1} - z_{ji} + u_{ji}^k||_2^2)$$

Above objective function is strictly convex, therefore solution is unique.Lets assume

$c_1 = \lambda * (1-\alpha) * w_{ij}, c_2 = \lambda * (\alpha) * w_{ij}, a = x_i^{k+1} + u_{ij}^k, b = x_j^{k+1} + u_{ji}^k$
There are 3 possible cases for the optimal values of $z_{ij}^* and z_{ji}^*$