MATH6380J Mini Project 1 Report

Yue JIANG (20238316)
Zhenzhen LI (20303719)
Lizhang MIAO (20296174)

March 13, 2017

1 Background and Data Description

In research field, academic society is a large scale and complex group. There are many people cooperating with other authors, citing other authors' paper, reading other people's paper everyday, which constructs a complicated and relatively active network. Data scientists are analysing big data everyday and also producing data in everyday research. Thanks to Prof. Jiashun Jin's data, which contains citation and coauthor data in statistics field (Ji et al., 2017).It covers all the papers between 2003 and the first quarter of 2012 from the Annals of Statistics, Journal of the American Statistical Association, Biometrika and Journal of the Royal Statistical Society Series B. The paper corrections and errata are not included. There are 3607 authors and 3248 papers in total. Researchers are looking for great papers highly related to their research field by themselves. However, as there are so many fresh papers coming out everyday sometime one may leave out some important and highly related papers. If one can be automatically recommended some papers that are significantly related to his research, it will surely help conduct research, especially for those very beginners. Our idea is to use Jin's data to build a recommendation system that when one is reading a paper, the system can adaptively recommend a few highly related papers to him.

2 Methodology

It is not hard to imagine if two papers have citation relationship, they might be related to each other. How to find out a few highly related paper is the main task. In standard binomial logistic model, one can regard present reading paper as a response $y_i \in \{+1, 0\}$, for $i = 1, 2, ..., n$ of covariates $x_i \in \mathbb{R}^p$, where x_i stands for i-th paper's citation column. In other words, if i-th paper has significant relationship with y_i,

$$\frac{P\{y_i = 1|x_i\}}{P\{y_i = 0|x_i\}} = exp\{\theta_0 + {x_i}^T\theta)\} \tag{1}$$

would be large. As the recommended papers are from a quantity and should not be so many, one should only be recommended by a few but highly related ones. Then it's natural to require θ not to be too dense, but relatively sparse.

By using the Linearized Bregman Algorithm to fit the sparse binomial logistic regression model in high dimension, one can computes the regularization path by minimizing log-likelihood of the binomial model under $\|.\|_1$ regularity, with the loss function as following

$$\mathbf{L}(\theta_0, \theta) = \frac{1}{n} \sum_{i=1}^{n} log(1 + exp(-y_i(\theta_0 + {x_i}^T \theta))) \tag{2}$$

Thanks to Libra R package, we can easily implement the algorithm.

3 Model Analysis

Here we use the paper-author relation matrix 3248×3607 size and assume one is reading the k-th paper, regarding it as y_k, and other 3247 columns as the variables x_i, the sparse binomial logistic model becomes:

$$min_{\theta^k} \frac{1}{n} \sum_{i=1, i\neq k}^{3248} \log(1 + exp(-y_k(i)(\theta_0^k + {x_i}^T \theta^k))) + \lambda \left\|\theta^k\right\|_1 . \tag{3}$$

From the estimated θ^k, we can tell that variables with positive coefficients will be those papers that the target reader is interested in. While those variables with negative coefficients will not be recommended to the target reader. Besides, from the regularized solution path of the optimization problem, we can further discern the relationship importance by looking at which variable is the first one that enters the solution path.
For example, if one is now reading the 100^{th} paper, "Exact calculations for false discovery proportion with application to least favorable configurations" written by *Etienne Roquain and Fanny Villers*, which papers will he be interested in among the remaining 3247 paper? Chances are that the reader will be interested in some papers written by the same group of authors. The penalized logistic regression results of the first paper against other papers is given in Figure.1. We can see that for most variables, or papers in this context, are zero, which means that these papers are not highly correlated to the target paper and they are written by other authors. There are two papers with positive coefficients with indices shown in Figure.1. Hence the probability of being interested in the 100^{th} paper will increase due to the existence of these two variables. So we would recommend these two papers to the reader. By checking the DOI's of papers, we find out that these two papers are "Goodness-of-fit tests for high-dimensional Gaussian linear models" written by *Nicolas Verzelen and Fanny Villers* and "Some nonasymptotic results on resampling in high dimension, II: Multiple tests" written by *Sylvain Arlot, Gilles Blanchard, and Etienne Roquain.* There are also several papers with negative coefficients with indices shown in Figure.1. So the probability of being interested in the 100^{th} paper will decrease to almost zero due to the existence of these papers. Similarly, we check the DOI's of these papers and find out that either these papers are written by different group of authors, or these papers are focusing on different fields from the 100^{th} paper. For example, the 193^{th} paper is "Adaptive estimation of stationary Gaussian fields" written by *Nicolas Verzelen.*

One more example, suppose the reader is reading the 99^{th} paper, "Multiple testing via FDRL for large-scale imaging data" written by *Chunming Zhang,*

Jianqing Fan, and Tao Yu. The first two recommendations made by the system are the 422^{th} paper, "Semiparametric detection of significant activation for brain fMRI" written by *Chunming Zhang and Tao Yu*, and the 1747^{th} paper "A Reexamination of Diffusion Estimators With Applications to Financial Model Validation" written by *Jianqing Fan and Chunming Zhang.*

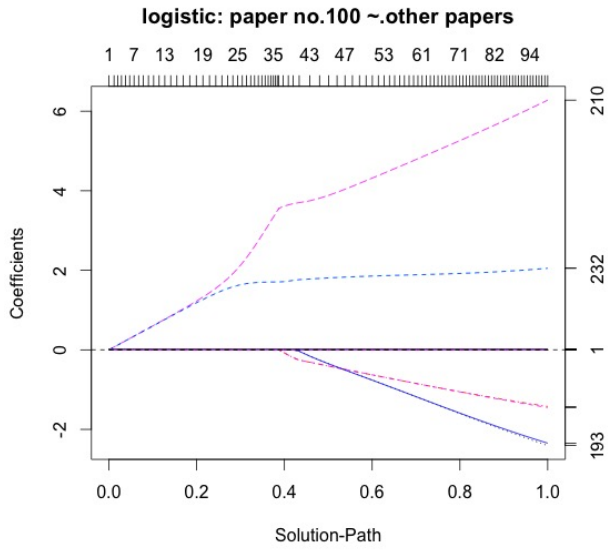

Figure 1: Regularized solution path for the 100-th paper reader

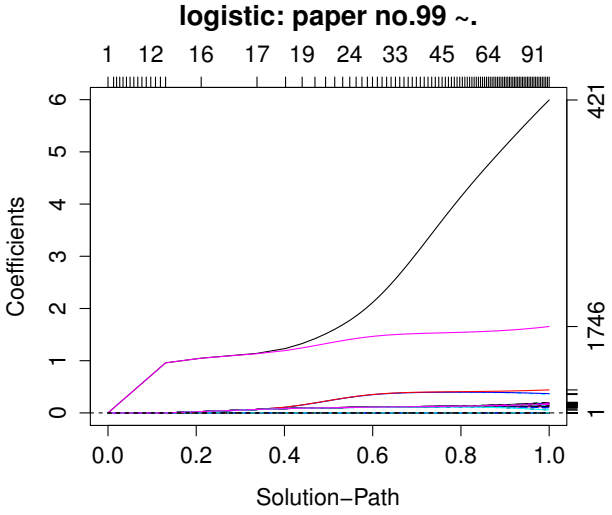

Figure 2: Regularized solution path for the 99-th paper reader

4 Summary

In this project, with data of citation and coauthor relationship in statistics field, benefiting from the Binomial Logistic Model under $\mathbf{L}_1$ regularity, the citation and coauthor network can be analyzed in various ways. Assume one is reading a paper at present, the log-likelihood of the binomial models gives highly related papers that one might be also interested in, and $\mathbf{L}_1$ gives a sparse solution that only a few paper would be recommended to him. With the author-paper network, imagine one is reading a paper at present, usually our model provides recommendation written by similar group of people as we have demonstrated in the implementation examples. Similarly, with the paper-paper citation network, one would be recommended by a few papers highly relation to the present ones, they might be citing the same papers or be cited by the same papers. In a word, those recommended papers are similar to some extent. One thing to point out is that the loss function assure that only similar papers would be chosen from and $\mathbf{L}_1$ penalty gives a sparse solution assuring that not too many papers would be recommended.

5 Remark

Group members' contribution:
Problem formulation: Yue Jiang, Zhenzhen Li, Lizhang Miao
Coding: Yue Jiang
Report writing: Yue Jiang, Zhenzhen Li, Lizhang Miao

6 Appendix

R code used in the project:

```
#1. get the ap matrix
ap<-read.table("authorPaperBiadj.txt",header=F) #author-paper matrix

#2. recommendation system
library(Libra)
y<-as.vector(2*as.matrix(ap[,100])-1)
X<-as.matrix(2*as.matrix(ap[,-100])-1)
path<-lb(X,y,kappa = 1,family='binomial',trate=100,normalize=FALSE)
plot(path,xtype='norm',omit.zeros=FALSE)
title(main=paste('logistic: paper no.100 ~.') ,line=3)
```

References

Pengsheng Ji, Jiashun Jin, et al. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2017.