CS229 Course Project Report

Multitask clustering of genes in different tissue types using Sparse PCA

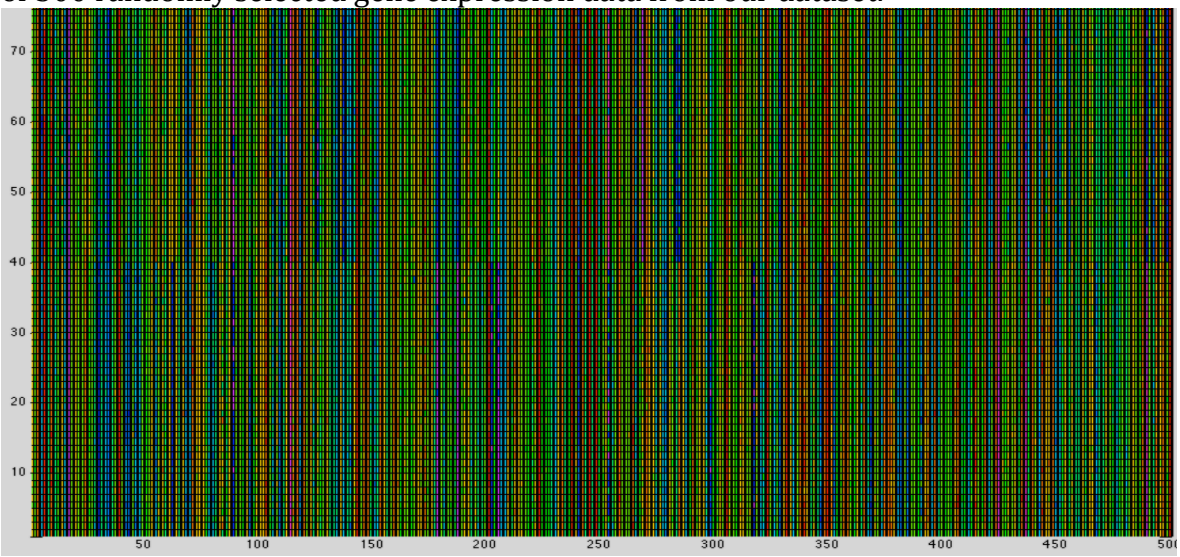
Margaret Zhang

I. Introduction

This project is a clustering application of gene expression data in different tissue types. We have a large dataset of gene expressions (78 samples*25134 gene sequences), where 39 samples are data from kidney and the other 39 from liver. The gene expression level of one specific gene sequence is quite similar in different tissue types. So our intuition is that most gene expression would be the same across all tissue types, yet cell differentiation lies in a small group of genes. We are particularly interested in identifying these gene sequences.

II. Motivation

The gene expression data are large, messy and highly correlated. Here is a plot of 500 randomly selected gene expression data from our dataset.



The color in each cell represents a certain level of gene expression activeness. From the figure we can see a clear boundary between sample 39 and sample 40, which indicates differences in gene expression activeness in different tissue types. However, the figure overall has great similarity among samples, which also verifies our intuition that most part of gene expression would be similar in cell differentiation. That's why we propose this method to identify similarity and difference in gene expression data of different tissue types.

Traditionally biological data are clustered using PCA technique. But results from PCA are also messy, meaning that loadings of each principle component or cluster are mostly non-zero. This property of PCA make interpretation extremely hard, as non-zero loadings would exhaust researchers' attention in finding significant gene sequence. So we propose using sparse PCA, a modification of traditional PCA to impose a similarity in clusters of different tissue types and a sparse loading of each cluster, which would help to solve this problem.

III. Problem Definition

The problem we want to tackle here is to build up a system that can identify similarities and differences among gene expressions in different tissue types. The hypothesis for this problem is that most gene expressions in different tissue types should be the same, while a small portion of genes expressed differently, which lead to different functionality of tissues.

In this project we only investigated into 2 tissue types for comparison. The input data of this system would be 2 n-by-p matrix, where n is the number of samples in a specific tissue type and p is the number of gene sequences extracted. Data in each cell of the matrix represents the gene expression activeness in the specific tissue type and specific sample. The higher the gene expression activeness is, more frequently that DNA sequence has been translated into proteins. The outputs are 2 clusters of 2 different tissue type, a shared loading matrix of the two clusters. And how the system works will be illustrated in the Approach section.

IV. Approach

A nice output from the system has the following properties: (1) 2 top K clusters from 2 different tissue types that are close enough but yet preserve some differences. (2) Shared loading matrix of the 2 cluster sets that is also sparse. (3) The model should also achieve best prediction error.

In order to achieve a nice output stated above, we use Sparse PCA technique, along with cross-validation to find the optimal parameters such as # of PCs and shrinkage parameters.

Sparse PCA

The idea of Sparse PCA is to solve the following objective function:

$$\min_{\mathbf{PC_1},\mathbf{PC_2},\mathbf{B}} ||\mathbf{X_1} - \mathbf{PC_1} * \mathbf{B}||_2^2 + ||\mathbf{X_2} - \mathbf{PC_2} * \mathbf{B}||_2^2 + \lambda_1 |\mathbf{B}|_1 + \lambda_2 |\mathbf{PC_1} - \mathbf{PC_2}|_1$$

We use X_1 and X_2 to denote the gene expression data for 2 different tissue types. Here PC_1 and PC_2 are the two cluster sets for X_1 and X_2. The first 2 terms of the objective function is to ensure minimum reconstruction error. We use L1 regularization for the last 2 terms, as sparsity would be nice property of L1 regularization. The term $\lambda_1|B|_1$ is to ensure a sparse shared loading matrix of the two clusters. And the last term is to ensure similarity between the two clusters constructed.

The way to solve this objective function is using the following algorithm:

Step 0: Assuming that the algorithm will produce PCs of size K, meaning that PC_1 and PC_2 are both n-by-K matrices. B would then be a K-by-p matrix. Initiate PC_1, PC_2 and B to be matrices with all entries zero.

Step 1: Assume PC_2 and B are fixed, the objective function becomes an ordinary LASSO of matrix PC_1 after a linear transformation. This can be solved for PC_1 with LARS algorithm that will be explained later.

Step 2: After getting an optimal PC_1 given PC_2 and B, we can do the similar procedure for PC_2 and B, which in turn will give optimal result given other 2 matrices constant.

Step 3: We can obtain the optimal results of PC_1, PC_2 and B by iterating between step 1 and step 2 until convergence.

LARS Algorithm to solve lasso problem:

General form of lasso problem: $\beta = \arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda|\beta|_1$

Step 0: Start with all coefficients β_j equal to zero.

Step 1: Find the predictor X_j most correlated with Y

Step 2: Increase the coefficient β_j in the direction of the sign of its correlation with y. Take residuals r=y-yhat along the way. Stop when some other predictor Xk has as much correlation with r as X_j has.

Step 3: Increase (β_j, β_k) in their joint least squares direction, until some other predictor X_m has as much correlation with the residual r.

Step 4: Continue until: all predictors are in the model

These two algorithms would give optimal results given parameters K, λ_1, λ_2. But in order to find the optimal parameters K, λ_1 and λ_2, we can try a grid of values of K, λ_1 and λ_2 and use cross-validation to find ones that minimize the prediction error.

Cross-Validation algorithm:

Step 1: For a given value of parameter set $\Theta = (K, \lambda_1, \lambda_2)$, divide the training set to be M roughly equal datasets. Use M-1 sets as training and the remaining one as a validation set.

Step 2: Use the M-1 sets as training to get the optimal results from the above objective function. Then use the remaining one set to compute prediction error. Do this for all combination of M-1 possible sets and compute the average prediction error. This would be the cross-validation error for the given parameter set Θ, called $CV(\Theta)$.

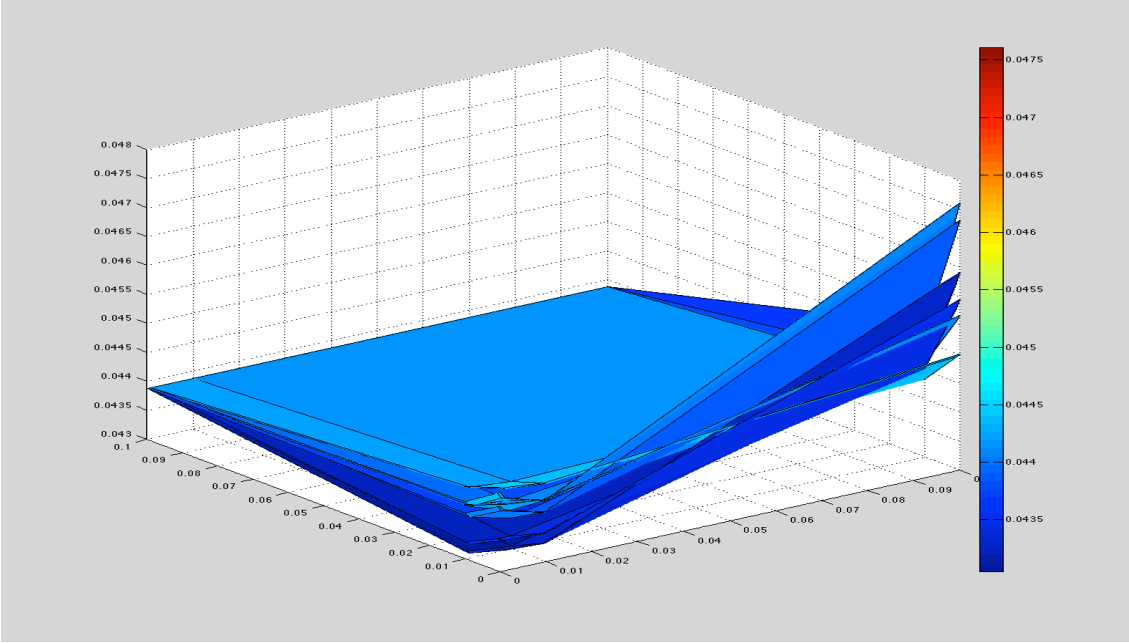
Step 3: Do step 2 for all possible values of Θ, and the final result would be Θ^* with minimum $CV(\Theta)$.

One last problem of building this system is how to evaluate the performance. For this system, the whole point is to find results that produce minimum prediction error. So prediction error is the primary evaluation of the system. The way to do this is to extract a small portion of the input dataset as the testing dataset, replace the holes with average of the specific gene sequence in specific tissue type. After that, using the new dataset as the input data and solve in the system. This test dataset would be used to evaluate performance of the system.

V. Result

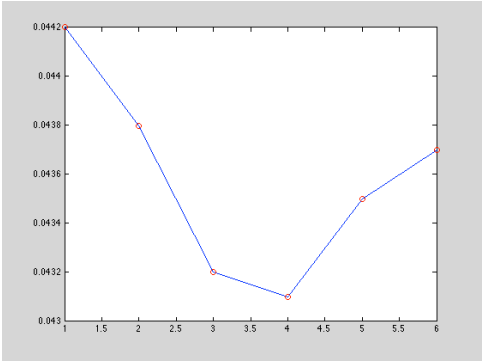
This result is important because it gives a guideline for scientists to investigate into genes that have different activeness level. And it may lead to other findings in genetic studies.

The cross-validation error for parameter set $\Theta = (K, \lambda_1, \lambda_2)$ has the following plots:



This plot clearly showed a surface of prediction errors of a grid of λ_1 and λ_2. Each surface is corresponding to a value K, the number of clusters produced by the system. All surfaces in the plot achieves the minimum cross-validation error at $\lambda_1 = 0.01$ and $\lambda_2 = 0.01$.

If given $\lambda_1 = 0.01$ and $\lambda_2 = 0.01$, the plot for different K values is:



Hence the value K that achieves the global minimum error is K = 4. We can then pick K = 4, $\lambda_1 = 0.01$ and $\lambda_2 = 0.01$ to be the optimal parameter set for this specific dataset.

Given these selected parameters, the system produces clusters of prediction error 0.0431 on the preselected test dataset.

VI. Future Work

After the system gives out two cluster sets for 2 tissue types, it is also critical to look into biological meaning of the results. Since we imposed for a similarity between the two cluster sets, it would be interesting to investigate into the difference between the two sets. That difference would potentially lead researchers to find gene sequences that played a significant role in cell differentiation.

VII. Reference

1. H. Zou, T. Hastie, R. Tibshirani. Sparse Principle Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265-286, 2006.
2. R. Tibshirani. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 58, No. 1, pages 267-288, 1996.
3. F. Mosteller and J.W. Tukey. Data analysis, including statistics. In *Handbook of Social Psychology*. Addison-Wesley, Reading, MA, 1968.