Gene Expression

Jeremy Meyer and Gabriel Adams

February 2019

1 Introduction

Genes are segments of DNA. They are essentially "code" for proteins found throughout the body. Different patterns of genes correspond to different actions and events including, importantly, cancer. Gene expressions are basically the number of genes used to perform a specific task. Gene expression profiling techniques involve analyzing these patterns found in genes. Our goal is to, using gene expression, to better understand cancerous tumors. Specifically, we hope to develop a method to determine which specific genes, if any, correspond to highly malignant tumors. Since doctors can "turn off" genes associated with malignant tumors, we hope these results can make cancer prevention and treatment more effective.

1.1 The data

Our data comes from a group of 102 patients. For each patient, measures for 5,049 different genes were recorded. These measures range from about 0 to just over 12. A "Malignant" score was also recorded for each patient. This score is a measure, ranging from 0 to 1, of the severity of a tumor, with a 1 indicating a very invasive tumor.

A small subset of the gene measures is shown in the Figure 1. It would be infeasible to show all of them. Through the plot, we can see relationships between the various gene measurements and the malignant score. It is worth noting that the Malignant score values tend to either be close to zero or 1 and some genes are very correlated with each other.

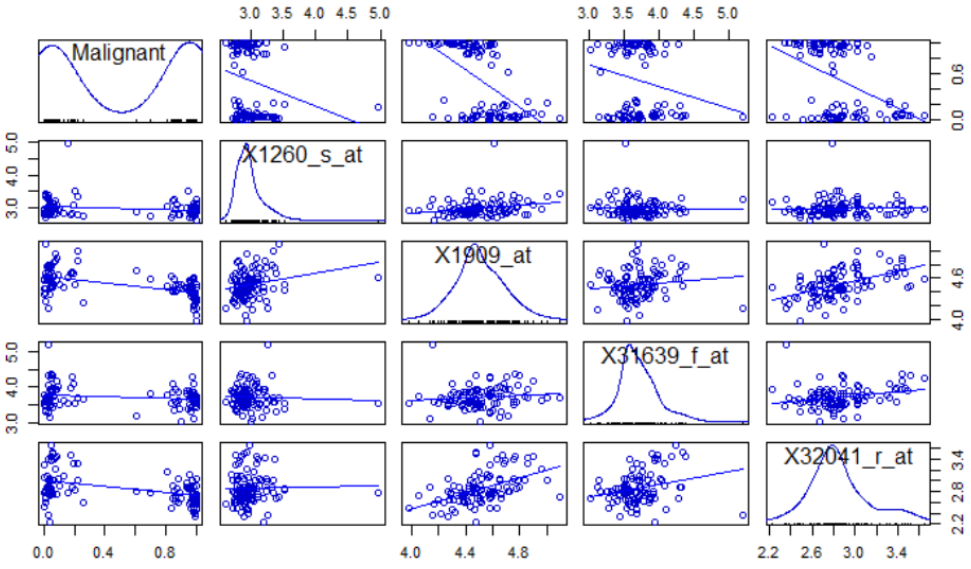

Figure 1: Pairs plots between Malignant Score and 4 sample genes

1.2 The Curse of Dimensionality

One problem that becomes apparent with the data is the sheer number of genes relative to the number of observations. If we tried fitting a multiple linear regression model with 5149 predictors and 102 data points, that would result in a system with infinitely many solutions. This is analogous to trying to find the best fitting line through a single point or a plane of best fit through 2 points. The line or plane could simply be rotated for infinitely many solutions. In fact, we could arbitrarily pick 102 of the genes and produce a model that perfectly predicts the Malignant score from our data.

Another potential problem that we have to be careful with in high dimensions is overfitting. As the number of dimensions of the data increases, the resulting volume of space grows so fast that the data points become very sparse. As a result, we don't have as much information about the space and our models can overreact to random noise in the data. This is known as the curse of dimensionality. One consequence of this is volatile model parameter estimates, which can cause problems when we make inference on which genes correspond to high malignant scores. Thus, if we are to use regression techniques, we must somehow reduce the dimensionality of the problem.

2 Model Selection

Due to the curse of dimensionality mentioned in the previous section, a constrained model will be necessary. We will use a penalized least squares technique. Instead of minimizing just the squared residuals (as in standard multiple linear regression), penalized least squares also minimizes coefficient size. This restricts the flexibility of our model and helps with the curse of dimensionality. It is important to note that while it will increase bias, it will greatly decrease the variance of our model. This reduction in variance means that when we add new data to our model, our estimates will not change as drastically as they would otherwise. We have few observations, so this is a very useful property. In addition, using penalized least squares will allow us too asses each gene for significance, which is necessary to answer our research questions.

We will use the elastic net method, which is a blend of LASSO and ridge regression. Ridge regression is useful because it performs well when there are lots of small coefficients (which we have with our gene data). LASSO is useful because it performs variable selection, which helps us determine which coefficients are truly important. However, when two explanatory variables are highly correlated, LASSO will somewhat arbitrarily remove one of them, even if the removed variable truly explains part of the response variable. Elastic net is a combination of both techniques, with α determining how much of each technique is used. If $\alpha = 1$, a LASSO model is fit. If $\alpha = 0$, a Ridge model is fit. If $0 < \alpha < 1$, then some combination of the two techniques is used. Using elastic net will help us balance the pros and cons of either method.

Elastic net seeks to find the $\boldsymbol{\beta}$ vector such that the following quantity is minimized:

$$\sum_{i=1}^{n}(y_i - x_i'\boldsymbol{\beta})^2 + \lambda \sum_{p=1}^{P}\left(\alpha|\beta_p| + (1-\alpha)\beta_p^2\right) \tag{1}$$

. Where each malignant score can be approximated by:

$$\boldsymbol{y} \approx \boldsymbol{X}\boldsymbol{\beta} \tag{2}$$

- y_i are the individual response variables (the values of Malignant). Together, they form the vector $\boldsymbol{y}$.
- x_i' is a vector of the various explanatory variables (the gene measurements) for each observation. Each is a row of the matrix $\boldsymbol{X}$. These were centered and scaled by their means and standard deviations so that the relative sizes of the gene measurements would be the same. This way, genes are not given more value due to their larger size. This will affect the interpretation of the β coefficients in the vector.
- $\boldsymbol{\beta}$ is a vector of constrained coefficients. The first element (β_0) corresponds to the intercept, while the following elements ($\beta_1, \beta_2, ...$) correspond to the effects of the various explanatory variables on the response. It is important to note that these effects are "constrained" by our use of elastic net. This means that they are slightly biased, but their variance is greatly decreased.

If all the gene measurements were at their mean, we would expect Malignant to be β_0, on average and subject to the constraint previously mentioned. Under constraint, as the gene measurement associated with β_1 increases by 1 standard deviation, Malignant will increase by β_1 on average.

- λ is called the shrinkage parameter. It is estimated using cross validation and controls how "strong" the penalties of lasso and ridge are. A higher λ means that coefficients will be shrunk much faster, potentially to zero (under LASSO), than with a lower value of λ.
- α is the previously mentioned parameter that controls how much ridge and LASSO are used. Again, if $\alpha = 1$, a LASSO model is fit. If $\alpha = 0$, a Ridge model is fit. If $0 < \alpha < 1$, then some combination of the two techniques is used.

3 Model Justification and Performance Analysis

To use elastic net, an assumption of linearity must be met. While it is infeasible to include a pairs plot of over 5,000 explanatory variables in this report, we can see in the pairs plot included in the introduction that linearity between the various explanatory variables and the response is reasonable. We will therefore assume linearity.

To use elastic net, we need to first determine values for α and λ. Again, α will control how much of LASSO and ridge are used. λ will determine how constrained our model is. To determine the value of α, we aim to minimize the mean square error of our model using a grid search of twenty values for α between 0 and 1. A plot of the α's and their corresponding MSEs is shown below in Figure 2. MSE is minimized when $\alpha = 0.85$ so we will use 0.85 as the alpha parameter in elastic net. That means that our model will based mostly on LASSO, but will still have a bit of ridge regression in it. Using $\alpha = 0.85$, we use cross validation to determine an optimal value of λ. A plot of $\log(\lambda)$ vs. mean-squared error is shown below in Figure 3. The mean-squared error is minimized when $\lambda = e^{-3.338} = 0.0355$.

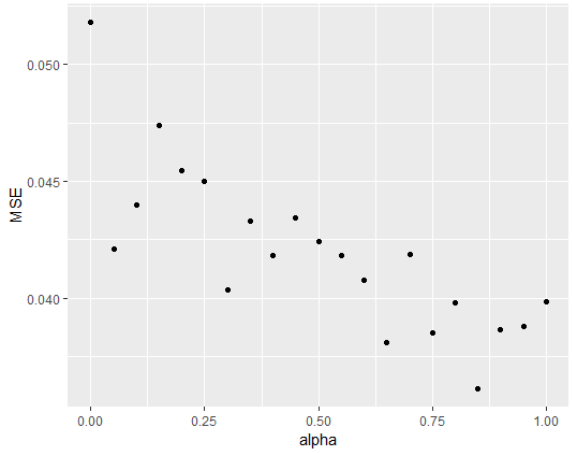

Figure 2

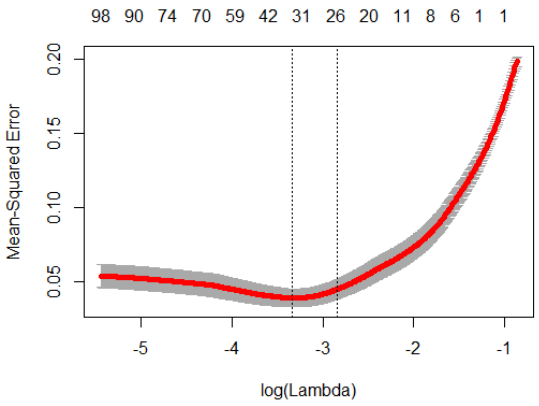

Figure 3

Using α and λ from above, we fit a elastic net model in R. We will explain the results derived from this model in the next section. Using this model, we determine a R^2 value of 0.9167. This means that about 91% of the variation in Malignant is explained by our model.

4 Results

After fitting our model in R, 36 variables remain in model. The others have been zeroed out by the LASSO component of the elastic net. Of these 36 variables, 10 are statistically significant. Their effects, as well as corresponding bootstrap confidence intervals, are included in Table 1.

Table 1: Genes that significantly affect Malignant

Gene	Estimate	Lower	Upper
X1260_s_at	-0.0689	-0.1379	-0.0042
X1909_at	-0.1899	-0.3797	-0.0341
X33716_at	-0.1861	-0.3721	-0.0497
X33921_at	-0.2030	-0.4060	-0.0958
X35331_at	0.1043	0.0139	0.2087
X37639_at	0.1615	0.1113	0.2570
X38087_s_at	-0.1297	-0.2318	-0.0978
X41584_at	-0.1258	-0.2517	-0.0083
X2041_i_at	-0.0723	-0.1446	-0.0138
X496_s_at	-0.0877	-0.1754	-0.0135

We can see that there are ten genes that have a significant effect on Malignant. Eight genes have a negative effect, while the other two have a positive effect. Holding all else constant and subject to the constraint, as X33921_at increases by 1 standard deviation, Malignant will decrease by 0.2030, on average, with a 95% confidence interval from 0.0958 to 0.2030. In other words, we are 95% confident that, holding all else constant and subject to constraint, as X33921_at increases by 1 standard deviation, Malignant will decrease by values between 0.0958 and 0.2030. Similar relationships hold for the other genes highlighted in blue above. Holding all else constant and subject to constraint, as X37639_at increases by 1 standard deviation, Malignant will increase by 0.1615, on average, with a 95% confidence interval from 0.1113 to 0.2570. A similar relationship holds for the gene X35331_at.

With the results from our table above, we are ready to answer the research question posed at the beginning of this report. The genes highlighted in red above have a positive effect on Malignant (as described in the previous paragraph). It might be wise to turn these genes "off" to reduce the risk of cancer. On the other hand, the genes highlighted in blue have a negative effect on Malignant (again, as described in the previous paragraph). Further research into these genes could be beneficial because they might help prevent cancer.

4.1 Comparison to other Methods

Another method we could have used to analyze the data and reduce the dimensionality of the problem is Principle Component Regression (PCR). PCR works by consolidating many variables into much fewer variables that still explain most of the variation in the data. A multiple regression model can then be performed on a vastly smaller set of variables that are uncorrelated. After back transforming to the original β coefficients, we were able to produce estimates comparable to the results obtained from the elastic net method. However, to help with the assumptions of the multiple regression linear model and extend the range of Malignant score to the real line, the malignant score was logit-transformed. The logit transform is defined as $logit(Y) = \log(Y/(1-Y))$ and preserves ordering since it is one-to-one and increasing. This means that an increase in malignant score on the transformed data corresponds to an increase in the original malignant score. We will refer to the transformed response as logitScore.

To protect against overfitting, we measured testing set error by cross validation. We used a total of 20 principle component variables because we did not gain much improvement in test set MSE by adding more components (see Figure 4).

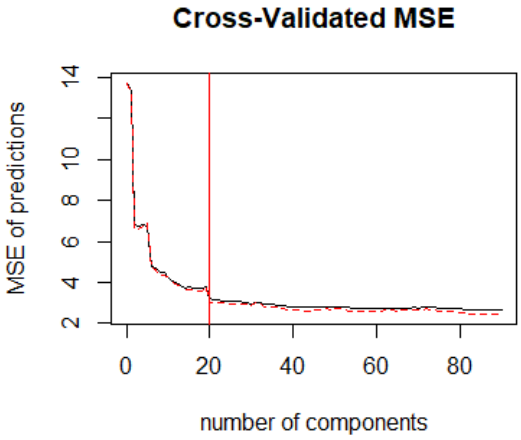

Figure 4: MSE from test set predictions for many different number of components.

Table 2: Top 10 genes that affected logit(score).

Gene	Beta_est	Std Err	2.5%	97.5%
X38218_at	0.0226	0.0025	0.0177	0.0276
X37639_at	0.0223	0.0014	0.0196	0.0250
X41106_at	0.0212	0.0026	0.0161	0.0264
X37124_i_at	-0.0211	0.0019	-0.0249	-0.0172
X37572_at	-0.0206	0.0022	-0.0249	-0.0163
X37347_at	-0.0197	0.0019	-0.0234	-0.0160
X38803_at	0.0191	0.0018	0.0156	0.0227
X2046_at	0.0185	0.0019	0.0147	0.0223
X41181_r_at	0.0184	0.0023	0.0139	0.0229
X34730_g_at	-0.0175	0.0013	-0.0200	-0.0150

The vast majority of gene effects were not zeroed out by PCR, but Table 2 shows the top 10 coefficients. Although these numbers are significantly lower than those in Table 1, there are substantially more genes still in the model and this are on the logit scale. Since the gene expression values were all centered and scaled before the analysis, the coefficients explain the expected effect one gene has on logitScore for one standard deviation increase. For example, we would expect that for each unit increase in standard deviation, gene X37639_at increases logitScore on average by 0.0223, holding all other gene expressions constant. The standard errors reflect our uncertainty due to sampling error and the 2.5% and 97.5% values represent bounds for 95% confidence intervals. If we were to resample from the population and generate more intervals, we would expect these to contain the true coefficient 95% of the time.

To determine if a gene effect is significant, we simply looked to see if 0 was contained in the interval. Based on this criteria, we found 2928 genes to have a statistically significant impact on malignant score! This is very different from the elastic net model results where only 10 genes were significant. It is worth noting that only one gene (X37639_at) appears significant in both methods. However, it also has the same sign (.16 compared to .02), which shows that the types of effects (positive or negative) are at least consistent for that gene. The R^2 value, or percent of variation explained by the PCR model relative to just using the mean of logitScore was 87.81%, which was slightly lower than that of the elastic net approach.

Telling doctors to look through 2928 genes may not be very practical. Although many of the genes could be slightly influential, it may be more useful to find a few genes are the *most* impactful. For this reason, we have decided to use elastic net as our final technique for determining which gene expressions have an effect on malignant score.

5 Conclusions

From the elastic net method, we were able to narrow a pool of 5149 genes into just 10 statistically significant genes by putting a constraint on the sizes of the coefficients. This allowed us to drastically reduce the dimensionality of the problem and identify genes that tend to affect highly malignant tumors (they are listed in Table 1). Since PCR left a substantial amount of significant coefficients and had a lower R^2 value, we decided to go with elastic net as our method of choice.

One major shortcoming to this method is simply the lack of data for each gene. It can be pretty difficult to separate effects of one gene amist over 5000 others with only 102 observations. The number of "significant" genes depends greatly on the model chosen, which can greatly affect our results. Even just the top 10 impactful genes for both PCR and elastic net ended up very different (only 1 gene was in both). Overall, these inferences should be treated as red flags for doctors to explore in a laboratory more carefully. If we had more than 102 observations, however, then we would be more confident in our inferences.

Moving forward, gathering more subjects will greatly enhance the power of our dimension reduction methods. It will give us more information about the high dimensional space, which will help reduce the variability in the estimated effects. With the results from this analysis, it may be helpful for doctors to carefully look at both the genes that tend to increase malignant scores and those genes that tend to result in lower scores. Doctors can avoid turning off genes that result in lower scores and study them in more depth. Experiments can also be done individually on genes that tend to have higher scores to try and isolate effects.

6 Teamwork Statement

Jeremy coded and completed the PC Regression part of the analysis. Jeremy also wrote the section on the curse of dimensionality, comparison to other methods, and conclusion. Gabe did the primary coding on the elastic net part of the analysis. Jeremy checked his work. Gabe also wrote much of the results, model justification and performance analysis, and model selection sections.