# Dimension Reduction and Regularization

Traditional machine learning methods start with the assumptions that the model features are (mostly) independent and that there are more samples then there are features (n > p). In our data set there are 47222 features (genes) and only 24 samples (patients), p is three orders of magnitude larger than n. In addition, genes express in groups, therefore many of the features are not independent. Attempting to use traditional methods such as least squares linear regression would result in severe overfitting and a poor predictive model on the Psoriasis dataset.

In this situation methods to reduce the dimensionality (discarding of “unnecessary” features) can offer a more reasonably sized model that may still accurately perform classification.

This paper examines three different methods for reducing a high dimensional data set like Psoriasis: Subset selection, Shrinkage, and Principal Component Analysis. The objective is to perform feature reduction and then assess the effectiveness by comparing the reduced dataset to the list of highly expressed genes from the Psoriasis paper. If the method is effective, then a many of the identified genes should still be present in the resulting, reduced model.

## Subset Selection

Three general approaches to subset selection are commonly used in statistical learning, with speed versus optimality the primary tradeoff. The results of these techniques can be rated using either cross validation or indirect training error estimates, such as RSS or R2

### Best Subset Selection:

The Best Subset method

Best Subset will provide an optimal model (based on either CV or indirect estimate) but at a significant processing cost. To reach the optimal model in the Psoriasis data set would require analysis of the almost inconceivably large 24722 potential models and is therefore not appropriate for our dataset.

Similarly, Backward Stepwise Selection, which would start from with the full, 47222 feature model, and works back removing the feature that provides the least value to the indirect estimate. Unfortunately, when the model has more features than samples, like ours, the full model will not fit and therefore cannot be used for our set.

That leaves Forward Stepwise Selection, which starts from the null model and adds the most useful feature at each stage. Forward selection requires inspection of on the order of 472222 potential models, not trivial but also not intractable. Because it starts with a model of no features it can be used when the full model has more predictors than samples.

*Step.r* performed forward stepwise reduction on the dataset using a logistic regression model (appropriate for binary classification) for fitting and produced reduced feature models based on the AIC criteria. The objective was to determine if a “good” reduced model would include the genes identified as important to the activation of T-cells in Psoriasis patients from Table 1 of the original paper.

Table XX: Forward Stepwise reduced model features

|  |  |  |  |
| --- | --- | --- | --- |
| Probe | Gene | Deviance | AIC |
| ILMN\_1683678 | SPATS2L | 0 | 4 |
| ILMN\_3286813 |  | 10.152 | 14.152 |
| ILMN\_3281502 |  | 11.93 | 15.93 |
| ILMN\_1735014 | KLF6 | 11.956 | 15.956 |
| ILMN\_1781285 | DUSP1 | 12.519 | 16.519 |
| ILMN\_1778617 | TAF9 | 12.753 | 16.753 |
| ILMN\_2321064 | BAX | 13.31 | 17.31 |
| ILMN\_3309534 |  | 13.765 | 17.765 |
| ILMN\_2143250 | FAR1 | 14.448 | 18.448 |
| ILMN\_3200597 |  | 14.904 | 18.904 |
| ILMN\_2119421 |  | 15.149 | 19.149 |
| ILMN\_1860954 |  | 15.579 | 19.579 |
| ILMN\_2285568 |  | 15.642 | 19.642 |
| ILMN\_2246956 | BCL2 | 15.764 | 19.764 |
| ILMN\_1731107 | CCDC92 | 15.922 | 19.922 |
| ILMN\_2397721 | GLB1 | 15.934 | 19.934 |

As Table XX illustrates, nine of the important genes are selected in the first seventeen features Forward Stepwise identified as the most useful from a minimal AIC standpoint, including SPATS2L and KLF6—the newly identified genes from the Psoriasis paper. The remaining genes appear in the first 108 features selected by the algorithm, implying that a model reduced from ~47000 to ~100 features would still include the primary genetic predictors.

## Shrinkage

The two shrinkage methods attempt to fit the full model through least squares (and attempting to minimize RSS) but with an additive penalty for larger coefficient values.

Ridge regression uses the easier to compute *l*2 penalty but the cost of this is a model that shrinks the magnitude of each coefficient but does not remove them. Figure XX is a plot of the ridge coefficients as a function of the shrinking penalty and demonstrated how the coefficients trend towards 0.

LASSO is a similar algorithm to Ridge but one that forces many coefficients to exactly zero—offering feature selection/reduction similar to Forward Stepwise. *Lasso.r* performs LASSO and Ridge on a logistic regression model fitted to the dataset across different shrinkage values. LASSO Models were selected manually for feature comparison to the Psoriasis paper list to determine if LASSO selected a similar set.

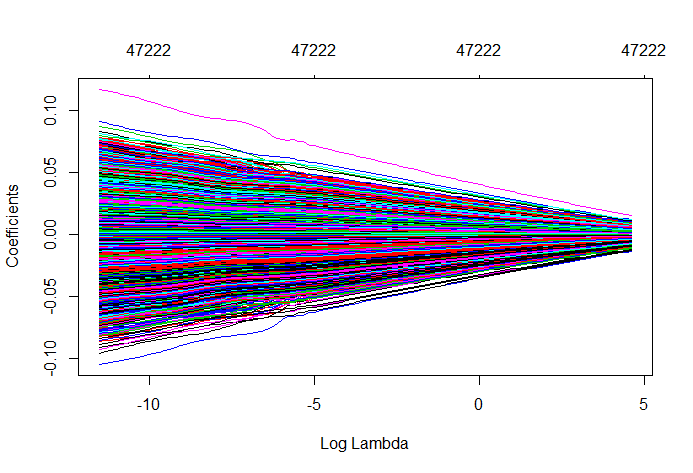
Figure XX: Ridge Regression

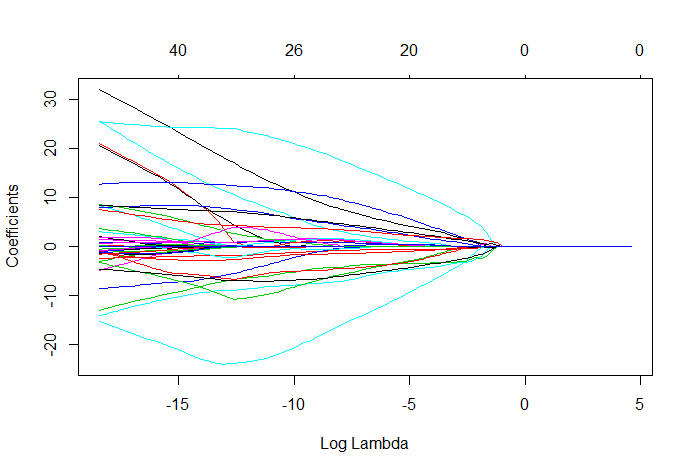
Figure XX: LASSO

Table XX: LASSO coefficients:

|  |  |  |
| --- | --- | --- |
| Probe | Gene | Beta(j) |
| ILMN\_1651913 |  | 21.15 |
| ILMN\_1652431 |  | -3.07 |
| ILMN\_1652784 |  | 0.78 |
| ILMN\_1654942 |  | 3.06 |
| ILMN\_1655827 |  | 1.40 |
| ILMN\_1656052 |  | 20.58 |
| ILMN\_1656421 |  | 0.09 |
| ILMN\_1659378 |  | -0.52 |
| ILMN\_1661886 |  | -1.10 |
| ILMN\_1662807 |  | 8.34 |
| ILMN\_1663767 |  | -4.82 |
| ILMN\_1667804 |  | 2.02 |
| ILMN\_1670219 |  | 0.27 |
| ILMN\_1670385 |  | 3.67 |
| ILMN\_1671004 |  | 0.01 |
| ILMN\_1673885 |  | 25.67 |
| ILMN\_1679647 |  | -0.02 |
| ILMN\_1683036 |  | -1.34 |
| ILMN\_1683678 | SPATS2L | -2.43 |
| ILMN\_1685540 |  | 8.64 |
| ILMN\_1688749 |  | -8.63 |
| ILMN\_1694742 |  | -14.19 |
| ILMN\_1703855 |  | 32.11 |
| ILMN\_1712913 |  | -3.31 |
| ILMN\_1722916 |  | -15.38 |
| ILMN\_1735014 | KLF6 | -0.76 |

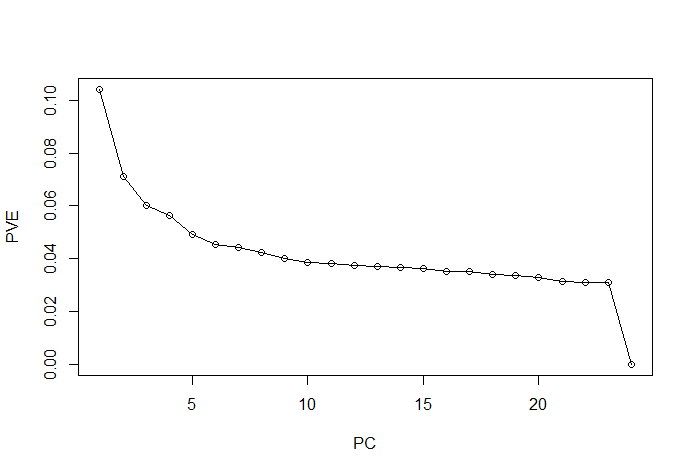
The variables selected by LASSO did include the two newly identified genes, SPATS2L and KLF6, but few of the other significantly up/down regulated genes listed. Our suspicion is the standardization of the expression levels masked some of the more strongly expressed genes by other associated genes. An interesting side note is the relatively moderate penalty required to force all the LASSO coefficients to zero. This is not surprising as the penalty is proportional to the sum of a function of the coefficients and with 47222 coefficients this summation will grow to a significant penalty quickly.

## Principal Component Analysis

The final reduction method explored is Principal Component Analysis. This method differs from the previous in that it doesn’t depend on a response to perform feature reduction. Instead it simply the “direction” along which the data is most varied. Subsequent components are found in the same manner with the added requirement that the direction of each new component much be orthogonal to the previous ones (on other words uncorrelated). The result is a reduced set of independent features each blended from the original data set. The downside of PCA versus subset or shrinkage methods is reduced visibility into the original feature set.

PCA.r performs the principal component analysis on the Psoriasis data set. As we are unable to compare the list of genes to that in the original paper the analysis how effectively the PCA based models predict the correct response. LOOCV validation was performed on models using the first 5, 10, 15 and 20 principal components fitted to a logistic regression model with the results displayed in Table XX.

Figure XX: PVE



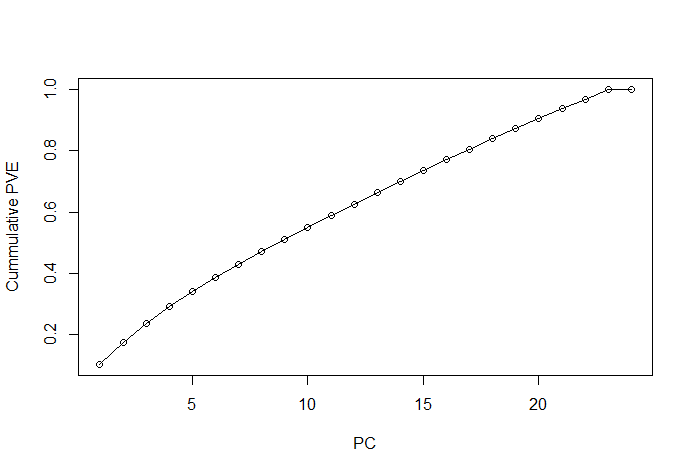
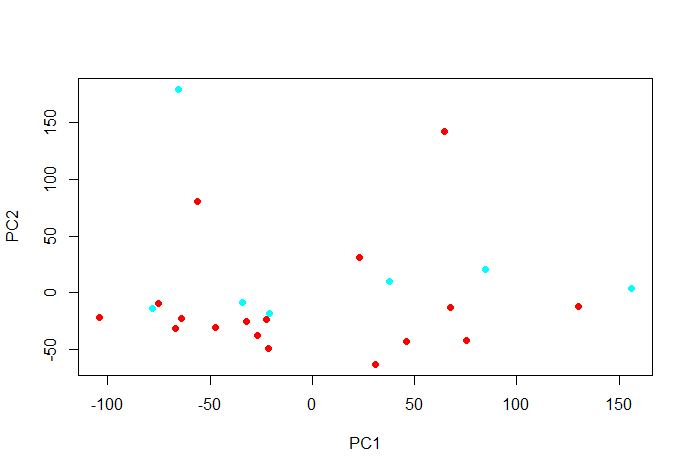


Figure XX: Visualizations of PCs (Red are patients, Blue are control samples)



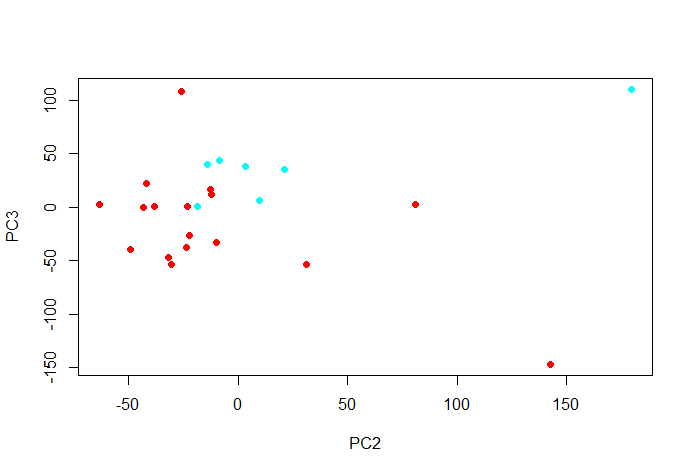


Table XX: Misclassification Rate for different sized PCA models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | All Samples | LOOCV | | | |
|  | Actual | 10 PCS | 5 PCs | 10 PCs | 15 PCs | 20 PCS |
| GSM1152973 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152974 | Yes | Yes | Yes | Yes | No | No |
| GSM1152975 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152976 | Yes | Yes | Yes | Yes | Yes | No |
| GSM1152977 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152978 | Yes | Yes | Yes | Yes | Yes | No |
| GSM1152979 | Yes | Yes | Yes | Yes | No | No |
| GSM1152980 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152981 | Yes | Yes | Yes | Yes | Yes | No |
| GSM1152982 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152983 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152984 | Yes | Yes | Yes | Yes | Yes | Yes |
| GSM1152985 | Yes | Yes | Yes | No | No | Yes |
| GSM1152986 | No | No | No | Yes | Yes | No |
| GSM1152987 | Yes | Yes | No | Yes | No | Yes |
| GSM1152988 | No | No | No | No | No | No |
| GSM1152989 | Yes | Yes | Yes | Yes | Yes | No |
| GSM1152990 | No | No | Yes | No | No | No |
| GSM1152991 | Yes | Yes | Yes | Yes | Yes | No |
| GSM1152992 | No | No | No | No | No | No |
| GSM1152993 | Yes | Yes | No | No | No | Yes |
| GSM1152994 | No | No | Yes | Yes | Yes | No |
| GSM1152995 | No | No | No | Yes | No | No |
| GSM1152996 | No | No | No | Yes | No | Yes |
| Error |  | 0 | 4 | 6 | 7 | 8 |
| Error Rate |  | 0.00% | 16.67% | 25.00% | 29.17% | 33.33% |

Figure XX illustrated the “Proportion of Variance Explained” by each principal component. The first component only explains ~10% of the variance with 100% of the variance reached with 23 principal components. Figure XX helps visualize the relationship between PCs, with the PC2/PC3 plot demonstrating good clustering between the two sample groups. Table XX illustrates an interesting result: as more principal components were added to the model, the ability of the model (using LOOCV validation) to accurately predict the presence of Psoriasis worsened. Examination the scree plot of PVE in Figure XX provides a clue to this result. At approximately seven principal components there is a “knee” were the additional amount of variance explained levels out. From this point on the additional principal components add little new information to the model in exchange for greatly increased noise and unwanted flexibility.