Statistical Methods For Dimension Reduction And Feature Selection For Integrating Genomic And Other Biological Data

Sande Sumaiya Zakirhusen

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali
April 2017

Certificate of Examination

This is to certify that the dissertation titled **“Statistical methods for dimension reduction and feature selection for integrating genomic and other biological data”** submitted by **Sande Sumaiya Zakirhusen** (Reg. No. MS12100) for the partial fulfillment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Shashi Bhushan Pandit

Dr. Lingaraj Sahu

Prof. Kapil H. Paranjape
(Supervisor)

Prof. Somdatta Sinha
(Supervisor)

Dated: April 18, 2017

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Prof. Somdatta Sinha at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of work done by me and all sources listed within have been detailed in the bibliography.

Sande Sumaiya Zakirhusen
(Candidate)

Dated: April 18, 2017

In my capacity as the supervisor of the candidates project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Prof. Somdatta Sinha
(Supervisor)

Prof. Kapil H. Paranjape
(Supervisor)

Acknowledgements

I would like to express my sincere gratitude to Prof. Somdatta Sinha for providing me the opportunity to work under her guidance and for giving me timely advice to understand my goals. I would also like to thank my external guide Dr. Samsiddhi Bhattacharjee for encouraging me to learn statistics and being helpful in all the cases. I am thankful to Prof. Kapil H. Paranjape for being my co-guide from Mathematics Department. I also express my gratitude to Dr. Lingaraj Sahu and Dr. Shashi Bhushan Pandit for giving the suggestions to improvise my work and evaluating me.

I also wish to thank Nitika Kandhari and friends in DBS who were always supportive to me. I wish to express sincere gratitude to Department of Mathematics, IISER Mohali and National Institute of Biomedical Genomics (NIBMG).

Sande Sumaiya Zakirhusen

List of Figures

Contents

Abstract

In recent times, many disciplines (like biology, chemistry or finance) have seen an explosion of data. The statistical methods face challenging tasks when dealing with such high-dimensional, multi-variate data. However, much of the data is highly redundant and can be efficiently brought down to a much smaller number of variables without a significant loss of information. The mathematical procedures making this reduction possible are called Dimensionality Reduction Techniques. Each and every technique reduces the dimensions of the data based on different criteria. This project has been done in three parts. In the first part, a simulation-based comparative study of variable selection was done in a linear-regression setting using a penalized-regression method - Least Absolute Selection and Shrinkage Operator (LASSO) versus univariate regression followed by the False Discovery Rate (FDR). Sensitivity, Specificity and Receiver Operating Characteristic (ROC) curves were used for comparison of these methods. In the second part, one of the Dimension Reduction Technique the Principal Component Analysis (PCA) was used to compare codon usage bias of HIV-1 viral genomes and genes to its human host using whole genome sequences. In the third part, Single Nucleotide Polymorphism (SNP) selection was done using Empirical Bayes strategy in Genome-wide Association Studies (GWAS).

Chapter 1

Introduction

During the last decade, life sciences have undergone a tremendous revolution with the accelerated development of high-throughput technologies and laboratory instrumentations. A good example is the biomedical domain that has experienced a huge advance since the advent of complete genome sequences. This post-genomic era has led to the development of new high-throughput techniques that are generating enormous amount of data, which has given rise to the exponential growth of many biological databases. In many cases, these datasets have much more variables than observations. Many scientific fields have seen an explosion of the number of variables measured for a single experiment. This is the case of image processing, mass spectrometry, time series analysis, internet search engines, etc. The statistical approach to study such data involves Multivariate Analysis [ZB86] methods.

1.1 Linear Regression

A linear regression model assumes that the regression function $f(X)$ is linear in the inputs $X_1, X_2, ...X_p$.. Given a vector of input variables $X = (X_1, X_2, ...X_p)$, if we want to predict a real-valued output Y, the linear regression model will have the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{1.1}$$

Here β_j's are unknown parameters. The linear model either assumes that the regression function $f(X)$ is linear, or that it is a reasonable approximation. The model is linear in parameters. If there are N number of data points (i.e., measurements) for each input variable X_j, for the response variable Y, the normal equations in matrix notation are:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & x_{N3} & \dots & x_{Np} \end{bmatrix} \tag{1.2}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad and\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \tag{1.3}$$

1.2 Feature Selection and Feature Extraction

Statistical and machine learning methods face a formidable problem when dealing with such high-dimensional data and normally the number of input variables is reduced before a data mining algorithm can be successfully applied. The dimensionality reduction can be made in two different ways: by only keeping the most relevant variables from the original datasets (*feature selection*) or by exploiting the redundancy of the input data and by finding a smaller set of new variables, each being a combination of the input variables, containing basically the same information as the input variables (*feature extraction*).

There are number of techniques available to reduce the dimensions of a large dataset. Each and every technique reduces the dimensions of the data based on different criteria. In recent years, Principal Component Analysis (PCA) [Hot33] and Linear Discriminant Analysis (LDA) [McL04] are regarded as the most fundamental and powerful tools of dimensionality reduction for extracting effective features of high-dimensional vectors in input data. Depending on the data, the reduction techniques are classified as linear techniques and non-linear techniques.

In this project, I have first learnt the mathematical insights behind different dimension reduction techniques that are used in Linear regression problems, such as the Least Absolute Selection and Shrinkage Operator (LASSO) [HTF09], False Discovery Rate (FDR) [Efr05], linear Principal Component Analysis (PCA), and then applied them on large sets of real biological data. In Chapter 2, I have shown, with a toy example, that LASSO performs better than FDR in most conditions. I have then studied two biological problems. In Chapter 3, I have compared the codon usage pattern of HIV-1 genes and its host human from 241 whole genome sequences using the PCA. The results show differential adaptation of codon usage in HIV-1 genes towards that of the human hosts. In Chapter 4, I have studied the Single Nucleotide Polymorphism (SNP) selection from Genome-wide Association Studies (GWAS) of the disease $Psoriasis$, where more than 5.8 lakh SNPs were analysed using empirical Bayes strategy in the logistic regression model. The results suggest that incorporating the empirical Bayes strategy certainly improves prediction of disease-associated SNPs.

Chapter 2

Comparison of Two Feature Selection Methods - LASSO and FDR

Before comparing the two feature selection methods, it is necessary to study basic method of ordinary least square to fit a linear regression model, its drawbacks and how to simplify the model with respect to mean-squared error by other methods, such as Subset Selection and Shrinkage methods [HTF09].

2.1 Ordinary Least Square Model

As per Section 1.1, let the best fit model be

$$f(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \ ; \ 1 \leq i \leq N$$

In this method we minimize the following Residual Sum of Squares (RSS):

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 && (2.1) \\ &= \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 && (2.2) \end{aligned}$$

From statistical point of view, this criterion is reasonable if the observations (x_i, y_i) represent the independent random draws from their population. Even if the x_i's not drawn randomly, the criterion is still valid if the y_i's are conditionally independent given the inputs x_i.

Using above matrices, we can write Residual Sum of Squares in the matrix form as follows:

$$RSS(\beta) = (Y - X\beta)^T(Y - X\beta) \tag{2.3}$$

This is a quadratic function in $p+1$ parameters. Differentiating with respect to β, we obtain

$$\begin{aligned} \frac{\partial RSS(\beta)}{\partial\beta} &= \frac{\partial RSS(\beta)}{\partial(Y - X\beta)}\frac{\partial(Y - X\beta)}{\partial\beta} \\ &= 2(Y - X\beta)^T(-X) \end{aligned}$$

(using if $y = x^T Ax$ then $\frac{\partial y}{\partial x} = x^T(A^T + A)$)

$$\begin{aligned} \frac{\partial RSS(\beta)}{\partial\beta^T\partial\beta} &= 2(-X^T)(-X) \\ &= 2X^T X \end{aligned}$$

Assuming (for a moment) that X has full column rank, and hence, $X^T X$ is positive definite, we set the first derivative to zero

$$\begin{aligned} 2(Y - X\beta)^T(-X) &= 0 && (2.4) \\ -2Y^T X + 2\beta^T X^T X &= 0 && (2.5) \\ (\beta^T X^T X)^T &= (Y^T X)^T && (2.6) \\ X^T X\hat{\beta} &= X^T Y && (2.7) \end{aligned}$$

$$\boxed{\hat{\beta} = (X^T X)^{-1}X^T Y} \tag{2.8}$$

$$\boxed{\hat{Y} = X\hat{\beta} = X(X^T X)^{-1}X^T Y} \tag{2.9}$$

From equation (2.7) we have,

$$X^T(Y - X\hat{\beta}) =< X, (Y - \hat{Y}) >= 0$$

which implies that $(Y - \hat{Y})$ is orthogonal to the column space of X that span the subspace of $\mathbb{R}^N$ and the resulting estimate $\hat{Y}$ is therefore the orthogonal projection of Y onto this subspace.

It might happen that the columns of X are not linearly independent, so that X is not of full rank. This would occur, for example, if two of the inputs were perfectly correlated (e.g. $x_2 = 3x_1$). Then X^TX is singular and the least square coefficients $\hat{\beta}$ are not uniquely defined. However $\hat(Y) = X\hat{\beta}$ are still the projection of Y onto the column space X; there is more than one way to express the projection in terms of the column vectors of X.

Assume that the observations y_i are uncorrelated and have constant variance σ^2 and that the x_i's are fixed (non random). Then

$$\begin{aligned}
var(\hat{\beta}) &= ((X^TX)^{-1}X^T)(\sigma^2 I)((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} \\
&= X^{-1}\sigma^2 I(x^T)^{-1} \\
&= (X^TX)^{-1}\sigma^2 \\
E(\hat{\beta}) &= (X^TX)^{-1}X^TE(Y) \\
&= (X^TX)^{-1}X^TE(X\beta) \\
&= (X^TX)^{-1}X^TX\beta \\
&= \beta
\end{aligned}$$

Hence $\hat{\beta}$ is an unbiased estimate of β.

2.2 Gauss-Markov Theorem

This theorem asserts that the least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates. Consider the mean squared error of an estimator $\tilde{\theta}$ in estimating θ:

$$\begin{aligned}
MSE(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 \\
&= Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2
\end{aligned}$$

The first term is the variance, while the second term is the squared bias. Mean squared error is intimately related to prediction accuracy. The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error of all linear estimators with no bias. However there may well exist a biased estimator with smaller mean squared error. Such an estimator would trade a little bias for a larger reduction in variance.

2.3 Subset Selection

In this approach, we retain only a subset of the variables, and eliminate the rest from the model. Least square estimate is used to estimate the coefficients of the inputs that are retained. The question of how to choose k inputs involves the tradeoff between bias and variance, and there are number of criteria that one may use. Typically we choose the model that minimises an estimate of the expected prediction error.

Forward step-wise selection starts with the intercept and then sequentially adds the predictor into the model that improves the fit most. Suppose our current model has k inputs, represented by parameter estimates $\hat{\beta}$ and we add in a predictor that results in estimates $\tilde{\beta}$, then the improvement in fit is often based on the F-statistic given by,

$$F = \frac{RSS(\hat{\beta}) - RSS(\tilde{\beta})}{RSS(\tilde{\beta})/(N-k-2)}$$

A typical strategy is to add in, sequentially, the predictor producing the largest value of F, and stopping when no predictor produces an F-ratio greater than the 90^{th} percentile of the $F_1, N-k-2$ distribution.

Backward Stepwise selection starts with the full model, and sequentially deletes predictors. Like Forward selection, it typically uses an F-ratio as above to choose the predictor to delete. In this case we drop the predictor producing the smallest value of F at each stage, stopping when each predictor in the model produces a value of F greater than the 90^{th} or 95^{th} percentile when dropped. Backward selection can only be used when $N \geq p$, while Forward Step-wise can be used always. There are hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the best move; these require a parameter to set the threshold between when an "add" move is chosen over a "drop" move.

2.4 Shrinkage Methods

By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model. However because it is a discrete process - variables are either retained or discarded - it often exhibits high variance, and therefore does not reduce the prediction error of the full model. Shrinkage methods are more continuous, and do not suffer as much from high variability.

2.4.1 Ridge Regression

Ridge regression shrinks the regression coefficients by imposing the penalty on their size. The ridge coefficients minimize a penalised residual sum of squares,

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left(\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{2.10}$$

Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage. An equivalent way to write the ridge problem is

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq s, \tag{2.11}$$

There is one-to-one correspondence between the parameters λ and s. When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. By imposing size constraint on the coefficients, this phenomenon is prevented from occurring.

The ridge solutions are not equivalent under scaling of the inputs, and so one normally standardises the inputs before solving. For reparametrization we use the centred inputs: each x_{ij} gets replaced by $x_{ij} - \bar{x}_j$. We estimate β_0 by $\bar{y} = \sum_1^N y_i/N$. The remaining coefficients get estimated using centred inputs.

Writing the equation (2.10) in matrix form,

$$RSS(\lambda) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \tag{2.12}$$

Differentiating with respect to β, we obtain

$$\begin{aligned}
\frac{\partial RSS(\beta)}{\partial\beta} &= \frac{\partial RSS(\beta)}{\partial(Y - X\beta)}\frac{\partial(Y - X\beta)}{\partial\beta} + 2\lambda\beta^T \\
&= 2(Y - X\beta)^T(-X) + 2\lambda\beta^T \\
\frac{\partial RSS(\beta)}{\partial\beta^T\partial\beta} &= 2(-X^T)(-X) + 2 \\
&= 2X^TX + 2
\end{aligned}$$

We set the first derivative to zero

$$\begin{aligned}
2(Y - X\beta)^T(-X) + 2\lambda\beta^T &= 0 \\
-2Y^TX + 2\beta^TX^TX + 2\lambda\beta^T &= 0 \\
(\beta^TX^TX + \lambda\beta^T)^T &= (Y^TX)^T \\
(X^TX + \lambda I)\hat{\beta} &= X^TY
\end{aligned}$$

$$\boxed{\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY} \tag{2.13}$$

where I is the $p \times p$ matrix. Ridge regression solves the problem of singularity even if X^TX is not of full rank.

2.4.2 The LASSO (Least Absolute Selection and Shrinkage Operator)

The LASSO estimate is defined by

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t, \tag{2.14}$$

Just as in Ridge regression, we can reparametrize the constant β_0 by standardizing the predictors; the solution for $\hat{\beta}_0$ is $\bar{y}$, and thereafter we fit a model without an intercept. The L_1 penalty makes the solutions non-linear in y_i and quadratic programming algorithm is used to compute them.

2.5 False Discovery Rate (FDR)

Suppose we have N null hypotheses to consider simultaneously, each with its own test statistic,

$$\begin{aligned}
\text{Null hypotheses: } & H_1, H_2, ..., H_i, ..., H_n \\
\text{Test statistic: } & z_1, z_2, ..., z_i, ..., z_n
\end{aligned}$$

N must be large for local FDR calculations, at least in the hundreds, but the z_i need not be independent. we assume that the N cases are divided into two classes, null or non-null, occurring with prior probabilities p_0 or $p_1 = 1 - p_0$, and with the density of test statistic z depending upon its class,

$$
\begin{aligned}
p_0 &= Pr(null) \\
p_1 &= Pr(non - null) \\
f_0(z) &: \textit{ density if null} \\
f_1(z) &: \textit{ density if non-null.}
\end{aligned}
$$

Define the null subdensity

$$f_0^+(z) = p_0 f_0(z)$$

and the mixture density

$$f(z) = p_0 f_0(z) + p_1 f_1(z)$$

The local False Discovery Rate is

$$
\begin{aligned}
FDR(z) &= Pr(null|z) \\
&= p_0 f_0(z)/f(z) \\
&= f_0^+(z)/f(z)
\end{aligned}
$$

2.6 Comparison of LASSO and FDR methods

Conventions:
X is a $n \times p$ data matrix, Y is $n \times 1$ response variable matrix. $\Sigma_{p \times p}$ covariance matrix of X. t is number of true β values. $\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$ for $1 \leq i \leq t$ and $\beta_i = 0$ for $(t+1) \leq i \leq p$.
The simulations are done for three cases: (a) $p \ll n$ (b) $p \sim n$ (c) $p \gg n$
And plots are produced for different values of a parameter π, where $\pi = t/p$.
For LASSO, 'glmnet' package (https://cran.r-project.org/web/packages/glmnet/index.html) is used while for FDR t-statistics are used (since FDR generally gives same rank of variables as the t-statistics) in 'R software' (www.r-project.org).

2.6.1 Performance criteria

LASSO and FDR methods were evaluated using the following measurements:

The True Positives (TP)

It is the number of positive decisions when tests are true or they can be called as correctly identified signals.

The True Negatives (TN)

It is the number of negative decisions when tests are false or they can be called as correctly rejected signals.

The False Positives (FP)

It is the number of positive decisions when tests are false or they can be called as incorrectly identified signals.

The False Negatives (FN)

It is the number of negative decisions when tests are true or they can be called as incorrectly rejected signals.

The True Positive Rate (TPR)

It is the probability that the decision is positive when the test is true. This is also called as Sensitivity and is defined by

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

The False Positive Rate (FPR)

It is the probability that the decision is positive when the test is false and is defined by

$$FPR = \frac{FP}{FP + TN}$$

Specificity

It is the probability that the decision is negative when the test is false and is defined by

$$\begin{aligned} Specificity &= 1 - FPR \\ &= \frac{TN}{TN + FP} \end{aligned}$$

Receiver Operating Curve (ROC)

It is created by plotting between TPR and FPR according to various cut-off values.

The Area Under ROC Curve (AUC)

It is used to measure the efficiency of a method.

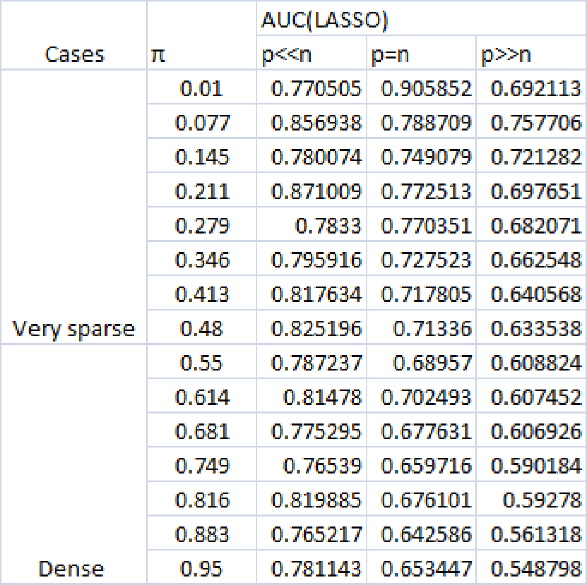
ΔAUC

This is defined as

$$\Delta AUC = AUC(LASSO) - AUC(FDR)$$

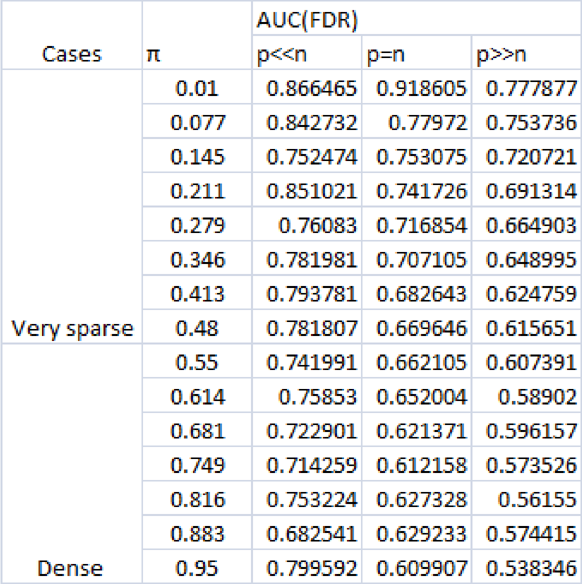
2.6.2 The simulation-based estimates

The simulation-based estimates of $AUC(LASSO)$, $AUC(FDR)$ and $\Delta AUC/AUC(LASSO)$ are as follows:



Cases	π	AUC(LASSO)		
		p<<n	p=n	p>>n
	0.01	0.770505	0.905852	0.692113
	0.077	0.856938	0.788709	0.757706
	0.145	0.780074	0.749079	0.721282
	0.211	0.871009	0.772513	0.697651
	0.279	0.7833	0.770351	0.682071
	0.346	0.795916	0.727523	0.662548
	0.413	0.817634	0.717805	0.640568
Very sparse	0.48	0.825196	0.71336	0.633538
	0.55	0.787237	0.68957	0.608824
	0.614	0.81478	0.702493	0.607452
	0.681	0.775295	0.677631	0.606926
	0.749	0.76539	0.659716	0.590184
	0.816	0.819885	0.676101	0.59278
	0.883	0.765217	0.642586	0.561318
Dense	0.95	0.781143	0.653447	0.548798

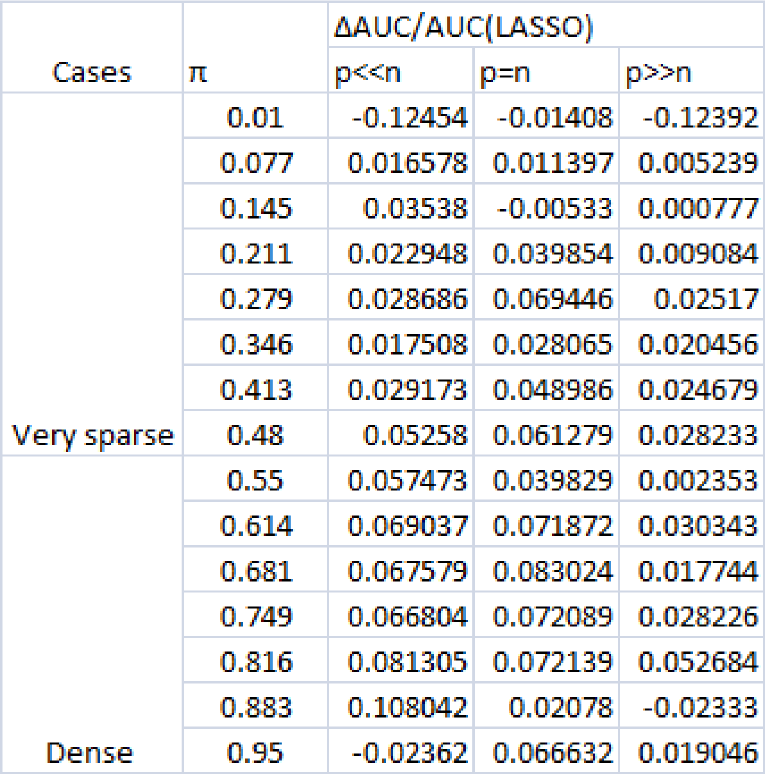
(a) AUC(LASSO)



Cases	π	AUC(FDR)		
		p<<n	p=n	p>>n
	0.01	0.866465	0.918605	0.777877
	0.077	0.842732	0.77972	0.753736
	0.145	0.752474	0.753075	0.720721
	0.211	0.851021	0.741726	0.691314
	0.279	0.76083	0.716854	0.664903
	0.346	0.781981	0.707105	0.648995
	0.413	0.793781	0.682643	0.624759
Very sparse	0.48	0.781807	0.669646	0.615651
	0.55	0.741991	0.662105	0.607391
	0.614	0.75853	0.652004	0.58902
	0.681	0.722901	0.621371	0.596157
	0.749	0.714259	0.612158	0.573526
	0.816	0.753224	0.627328	0.56155
	0.883	0.682541	0.629233	0.574415
Dense	0.95	0.799592	0.609907	0.538346

(b) AUC(FDR)

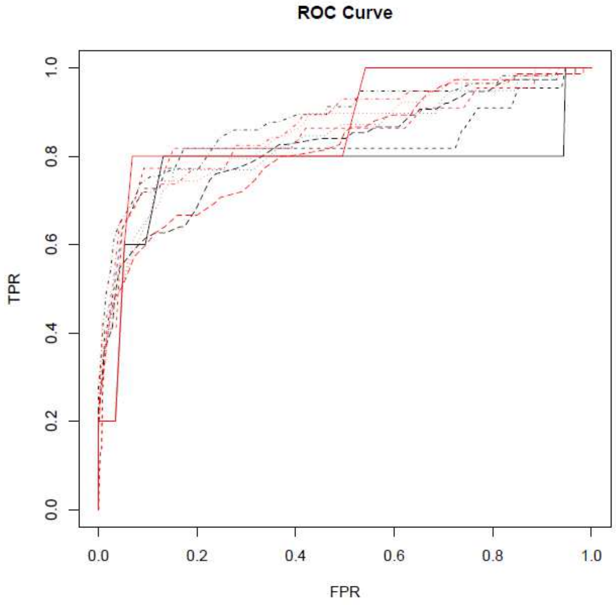
Figure 2.1: Area under ROC of both the methods LASSO and FDR, for different values of π for the three cases - $p \ll n$, $p = n$, and $p \gg n$ for $p = 2000$ and varying n as 500, 2000 and 4000



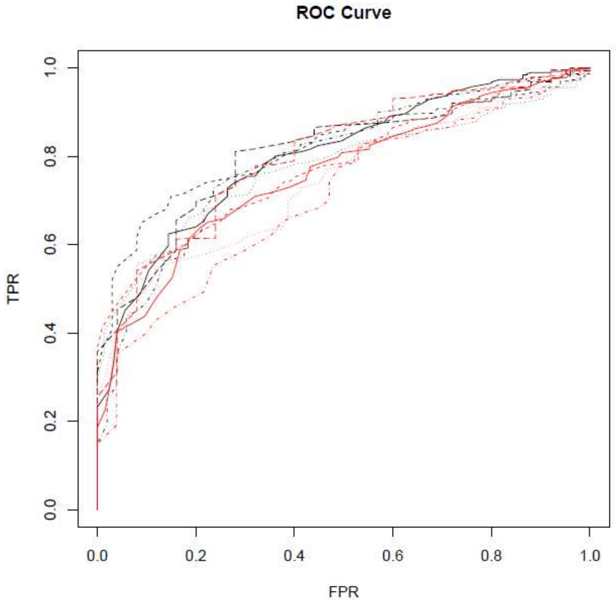
Cases	π	ΔAUC/AUC(LASSO)		
		p<<n	p=n	p>>n
Very sparse	0.01	-0.12454	-0.01408	-0.12392
	0.077	0.016578	0.011397	0.005239
	0.145	0.03538	-0.00533	0.000777
	0.211	0.022948	0.039854	0.009084
	0.279	0.028686	0.069446	0.02517
	0.346	0.017508	0.028065	0.020456
	0.413	0.029173	0.048986	0.024679
	0.48	0.05258	0.061279	0.028233
Dense	0.55	0.057473	0.039829	0.002353
	0.614	0.069037	0.071872	0.030343
	0.681	0.067579	0.083024	0.017744
	0.749	0.066804	0.072089	0.028226
	0.816	0.081305	0.072139	0.052684
	0.883	0.108042	0.02078	-0.02333
	0.95	-0.02362	0.066632	0.019046

Figure 2.2: Relative Difference = $\Delta AUC/AUC(LASSO)$

2.6.3 ROC of LASSO and FDR methods

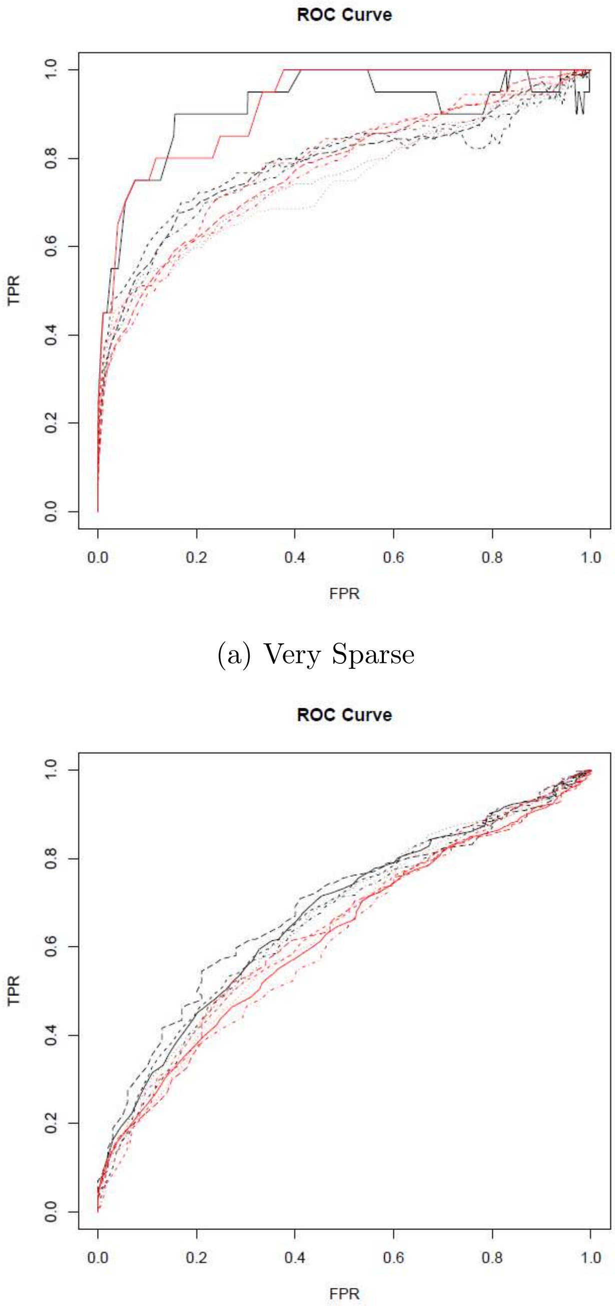


(a) Very Sparse



(b) Dense

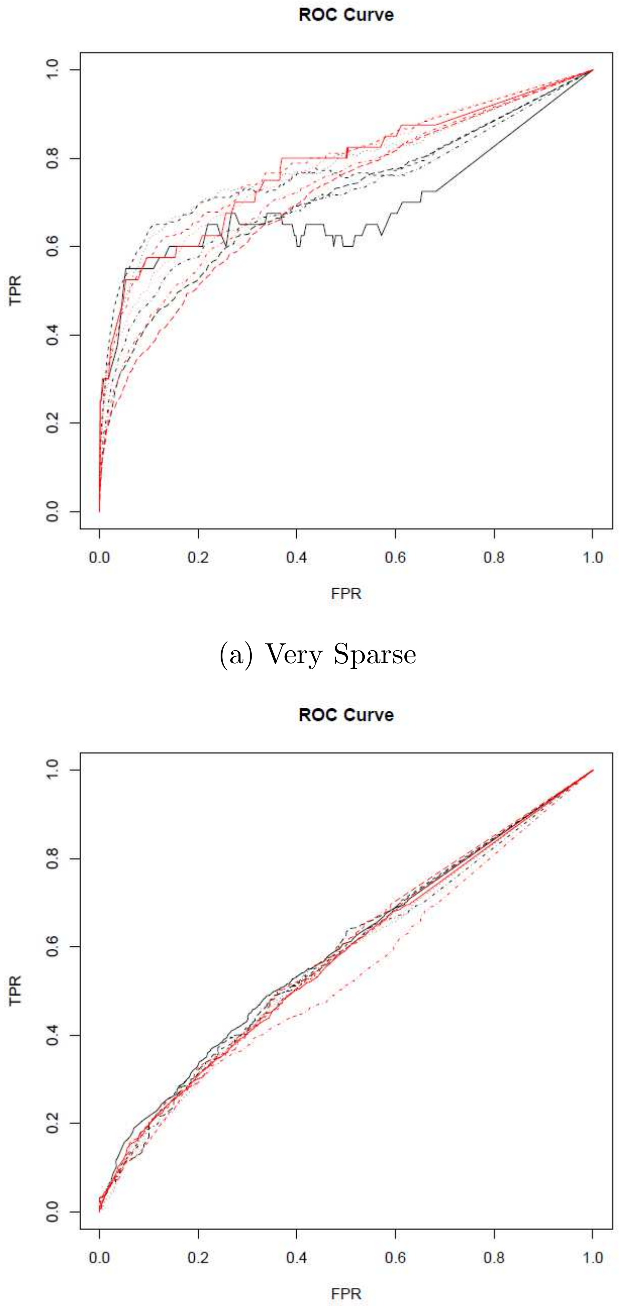
Figure 2.3: ROC (TPR vs FPR) for $p \ll n$ case. Red lines represent FDR method and Black lines represent LASSO method. (a) The ROC curves for π values $\pi = 0.01$ (solid line), $\pi = 0.045$ (dashed line), $\pi = 0.080$ (dotted line), $\pi = 0.115$ (dot-dash line), $\pi = 0.15$ (big dashed line). (b) The ROC curves for π values $\pi = 0.75$ (solid line), $\pi = 0.80$ (dashed line), $\pi = 0.85$ (dotted line), $\pi = 0.90$ (dot-dash line), $\pi = 0.95$ (big dashed line)



(a) Very Sparse

(b) Dense

Figure 2.4: ROC (TPR vs FPR) for $p = n$ case.
Red lines represent FDR method and Black lines represent LASSO method. (a) The ROC curves for π values $\pi = 0.01$ (solid line), $\pi = 0.045$ (dashed line), $\pi = 0.080$ (dotted line), $\pi = 0.115$ (dot-dash line), $\pi = 0.15$ (big dashed line). (b) The ROC curves for π values $\pi = 0.75$ (solid line), $\pi = 0.80$ (dashed line), $\pi = 0.85$ (dotted line), $\pi = 0.90$ (dot-dash line), $\pi = 0.95$ (big dashed line)



(a) Very Sparse

(b) Dense

Figure 2.5: ROC (TPR vs FPR) for $p \gg n$ case
Red lines represent FDR method and Black lines represent LASSO method. (a) The ROC curves for π values $\pi = 0.01$ (solid line), $\pi = 0.045$ (dashed line), $\pi = 0.080$ (dotted line), $\pi = 0.115$ (dot-dash line), $\pi = 0.15$ (big dashed line). (b) the ROC curves for π values $\pi = 0.75$ (solid line), $\pi = 0.80$ (dashed line), $\pi = 0.85$ (dotted line), $\pi = 0.90$ (dot-dash line), $\pi = 0.95$ (big dashed line)

2.7 Observations

▶ When $p \ll n$, $\Delta AUC/AUC(LASSO)$ increases as truths (π) increases.

▶ $AUC(LASSO)$ and $AUC(FDR)$ decreases as truths increases for $p = n$ and $p \gg n$ cases.

▶ $AUC(LASSO)$ and $AUC(FDR)$ decrease as we go from $p \lll n$ to $p \ggg n$ for almost all the truths (π).

▶ For most cases, ΔAUC is positive.

2.8 Conclusion

In general LASSO performs better than univariate t-test, except in highly sparse cases. The reasons for above observations of AUC pattern will be investigated in the future.

Chapter 3

Principal Component Analysis and its Application to Codon Usage Patterns

3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the feature extraction methods to identify patterns in data, and expressing the data in such a way as to highlight their similarities and differences. One of the main advantage of PCA is that once these patterns are found in the data, the data can be compressed (i.e. the number of dimensions can be reduced) without much loss of information. This method also solves the problem of correlation among the variables.

3.1.1 Method

Step 1 : Get a dataset

Let R be a data matrix of dimension $n \times p$ where there are p number of variables (columns) and n number of observations (rows).

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & r_{22} & r_{23} & \dots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & r_{np} \end{bmatrix} \tag{3.1}$$

Step 2 : Normalize the dataset

Variables can be in different units or scales. So, we need to normalize them to make them comparable. Normalized matrix will be

$$X = \begin{bmatrix} \frac{r_{11}-\bar{r_1}}{\sigma_{r_1}} & \frac{r_{12}-\bar{r_2}}{\sigma_{r_2}} & \frac{r_{13}-\bar{r_3}}{\sigma_{r_3}} & \dots & \frac{r_{1p}-\bar{r_p}}{\sigma_{r_p}} \\ \frac{r_{21}-\bar{r_1}}{\sigma_{r_1}} & \frac{r_{22}-\bar{r_2}}{\sigma_{r_2}} & \frac{r_{23}-\bar{r_3}}{\sigma_{r_3}} & \dots & \frac{r_{2p}-\bar{r_p}}{\sigma_{r_p}} \\ \multicolumn{5}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} \\ \frac{r_{n1}-\bar{r_1}}{\sigma_{r_1}} & \frac{r_{n2}-\bar{r_2}}{\sigma_{r_2}} & \frac{r_{n3}-\bar{r_3}}{\sigma_{r_3}} & \dots & \frac{r_{np}-\bar{r_p}}{\sigma_{r_p}} \end{bmatrix} \tag{3.2}$$

where, r_j be the j^{th} column of data matrix R, $\bar{r_j} = \frac{\sum_i r_{ij}}{n}$, and σ_{r_j} is the standard deviation of r_j^{th} column or variable..
Now let,

$$x_{ij} = \frac{r_{ij} - \bar{r_j}}{\sigma_{r_j}} \tag{3.3}$$

The normalized data matrix will be

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \multicolumn{5}{c}{\dots\dots\dots\dots\dots\dots\dots\dots} \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \tag{3.4}$$

Step 3 : Calculate the covariance matrix

The covariance matrix of $X_{n\times p}$ is given by

$$C_X = \frac{1}{n-1}X^TX = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \dots & \sigma_{1p}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 & \dots & \sigma_{2p}^2 \\ \multicolumn{5}{c}{\dots\dots\dots\dots\dots\dots\dots} \\ \sigma_{p1}^2 & \sigma_{p2}^2 & \sigma_{p3}^2 & \dots & \sigma_{pp}^2 \end{bmatrix} \tag{3.5}$$

where σ_{ij}^2 is the covariance between i^{th} and j^{th} variables (columns).

Step 4 : Calculate eigen values and eigen vectors of covariance matrix

According to the Singular Value Decomposition theorem, for a $m \times n$ matrix A whose entries are real or complex numbers, there exists a factoriazation (Singular Value Decomposition) of the form

$$A = U\Sigma V^{\star} \tag{3.6}$$

where, U is a $m \times m$ unitary matrix, Σ is a diagonal $m \times n$ matrix with diagonal entries as singular values, V is a $n \times n$ unitary matrix and $V^\star$ is the conjugate transpose of V and thus also unitary.
Hence, data set X can be written as a singular value decomposition as

$$\begin{aligned}
X &= U\Sigma V^\star \\
X^T X &= (U\Sigma V^\star)^T (U\Sigma V^\star) \\
(X^T X)V &= V\Sigma^2 V^\star V \\
(X^T X)V &= V\Sigma^2
\end{aligned}$$

This reduces to a eigen value problem, where V is the matrix of eigen vectors of $X^T X$ and Σ^2 is the square matrix whose diagonal entries are the eigen values of $X^T X$. Hence,

$$C_X V = \frac{1}{n-1}(X^T X)V = \frac{1}{n-1} V\Sigma^2 \tag{3.7}$$

These eigen vectors, when arranged in decreasing order of eigen values, form the new directions or dimensions.

Step 5 : Deriving the new data set

The above analysis gives the new rotated directions. Now, we need to project the original data on these new rotated axes. Hence, the new dataset is

$$Z = XV \tag{3.8}$$

Z is called as **Score Matrix**. The columns Z_i's of Z are called the **Principal Components** (PC). They account for the variance in the data in decreasing order. Here, the i^{th} Principal Component is given by

$$PC_i = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_p \end{bmatrix} \begin{bmatrix} v_{1i} \\ v_{2i} \\ v_{3i} \\ \vdots \\ v_{pi} \end{bmatrix} \tag{3.9}$$

where, $i \in [1, n]$. Variance accounted for i^{th} PC is given by

$$\begin{aligned} Var(PC_i) &= Var(Xv_i) \\ &= v_i^T Cov(X) v_i \\ &= v_i \frac{V\Sigma^2 V^\star}{n-1} v_i \\ &= \frac{(\Sigma^2)_{ii}}{n-1} \end{aligned}$$

Hence, when eigen values are arranged in the decreasing order, i^{th} eigen value represents the variance accounted by i^{th} Principal Component.

3.2 Application of PCA to Codon Usage Data of Human and HIV-1 Genes

3.2.1 Introduction

The information about the making of an organism is transferred between generations through its DNA, which is composed of a string of four nucleotide bases A, T, G and C. The information coded in this string is elaborated through transcription (DNA to mRNA) and translation (mRNA to protein) which function in the regulation of the biochemical pathways and other cellular processes. Many viruses store their genetic information in RNA (*Retroviruses*), which reverse transcribe the RNA to DNA, integrate their DNA into the host cellular DNA, and then complete the protein synthesis process. The Human Immunodeficiency Virus, HIV-1, which is the causative agent of the disease AIDS, is a RNA virus. On entering the host cells in human, it integrates to the host DNA. HIV-1 uses the host's translational machinery to translate their own genes. Thus, HIV-1 can be called as 'a translational parasite' on humans. Additionally, HIV genome is rich in A and T nucleotides, whereas human genome is G-C rich. This implies that HIV genes are translated poorly in the G-C rich host. Thus, deciphering the molecular basis of host-pathogen interactions is useful in understanding the factors governing pathogen evolution and disease propagation.

The three-letter genetic code used for translation from mRNA to proteins is degenerate, as multiple codons can code for the same amino acid (synonymous codons). Most organisms exhibit differences in base composition and significant codon bias (unequal usage of synonymous codons). Synonymous mutations (that alter the code but not amino acid) can change the base composition of genes without altering the corresponding protein sequence. Intuitively, synonymous mutations appear to be 'neutral' or 'near neutral' in their effects, however, their evolutionary consequences are being increasingly understood [PRL04, EP04, UH03, A08]. Studies show that codon bias and synonymous mutations are under weak selection, driving evolution in various organisms [CPH06, Ike85, Sto08]. Genes that are enriched for the preferred codons are known to have higher translational efficiency [CH09, KCT+03]. It has been shown in few bacteria-bacteriophage host-pathogen systems that long term co-evolution has resulted in some genes of bacteriophage being enriched in the codons preferred by their respective bacterial hosts [LNKP08]. In contrast, several RNA viruses show low association with their hosts in both base composition and codon usage [JH03]. Rationally, for a pathogen, which translates inside the host, having a codon bias similar to host's bias would help it to exploit the host's resources more efficiently. A balance between selection, mutation, and genetic drift maintains codon bias in host and pathogens. Thus, studies revealing determinants of the bias and its dynamics are central to the understanding of host-pathogen evolution.

In this project, we examine the codon usage pattern as a signature of pathogen genome evolution in host-pathogen interactions, for the Human Immunodeficiency Virus type 1 (HIV-1), in relation to its human host. HIV-1 is capable of undergoing rapid sequence level changes to evade the host's defence machinery, which is partially a result of strong immune selection pressure levied by the host. HIV-1 genome comprises of nine genes, which can be divided into two classes based on their function: structural genes (*env*, *gag* and *pol*) that form the essential components of the virus particle [Wil03, WPB+05]; and regulatory genes (*nef*, *rev*, *tat*, *vif* and *vpu*). Of these, regulatory genes *rev* and *tat* are mandatory [AC91, DSR+86, FFJ+86] and the remaining four genes are termed "accessory", as they are required for efficient replication of virus, but have been shown to be mandatory for survival in vitro [AH04, MS97, Tro95]. Below I show the codon usage pattern for the nine HIV-1 genes and compare them to that of the human genes.

3.2.2 Materials and Methods

Data Acquisition

To perform large scale codon usage analysis of the nine genes of HIV-1, we chose 241 whole genomes collected over a period of 13 years. HIV-1 genome sequence files were downloaded (in FASTA format) from the Los Alamos National Laboratory HIV

Sequence Database (www.hiv.lanl.gov, Oct 2016). 241 whole genome sequences of various subtypes deposited from year 1983 to 1995 were analysed. To avoid database bias, several checks were carried out. Only one sequence per patient was downloaded per year in order to avoid bias within a year. To avoid bias in gene-specific sequence deposition, we used Gene Cutter (www.hiv.lanl.gov) to clip all the genes from the whole genome sequence. Thus, equal number of sequences for all genes were available for each year, which were used for the analysis. Codon usage table for the average of human genes was retrieved from the Codon Usage Table Database (http://www.kazusa.or.jp/codon/) derived from GenBank Release 160.0 (June 15, 2007) [NGI00].

Codon Usage Analysis

Different scaling methods have been proposed for the study of codon usage pattern as they are associated with different biases like gene length, amino acid composition and the number of synonymous codons [PT02]. In order to avoid such biases in the codon usage data, I used the normalized codon frequency values calculated by scaling the frequencies for each codon with respect to the maximally-used synonymous codon for a particular amino acid [SST05],

$$\boxed{n_{ij} = \frac{x_{ij}}{x_{jmax}}} \tag{3.10}$$

where, n_{ij} is the normalized value for i^{th} codon and j^{th} amino acid, x_{ij} is the frequency of i^{th} codon for the j^{th} amino acid, and x_{jmax} is the frequency of the maximally used synonymous codon for the j^{th} amino acid. The start codon (AUG), the single codon (UGG) for tryptophan, and the three stop codons (UAA, UAG and UGA) were not included in the analysis.

Statistical Analysis

Multivariate Analysis methods are performed to reduce large number of variables or dimensions to a smaller number of new variables, so that they can be analysed while preserving most of the information of the original data. PCA was performed on the codon usage data for the nine HIV-1 genes and the average of human genes. Here we had 59 variables corresponding to the 59 degenerate codons, with 10 observations corresponding to the normalized codon usage value for HIV-1 genes and human. Thus the total data analysed is (241 genomes × 9 genes × 59 codons) plus the average codon usage of the human genes. Data curation was done manually, and the 'R software' (www.r-project.org) was used for the analysis.

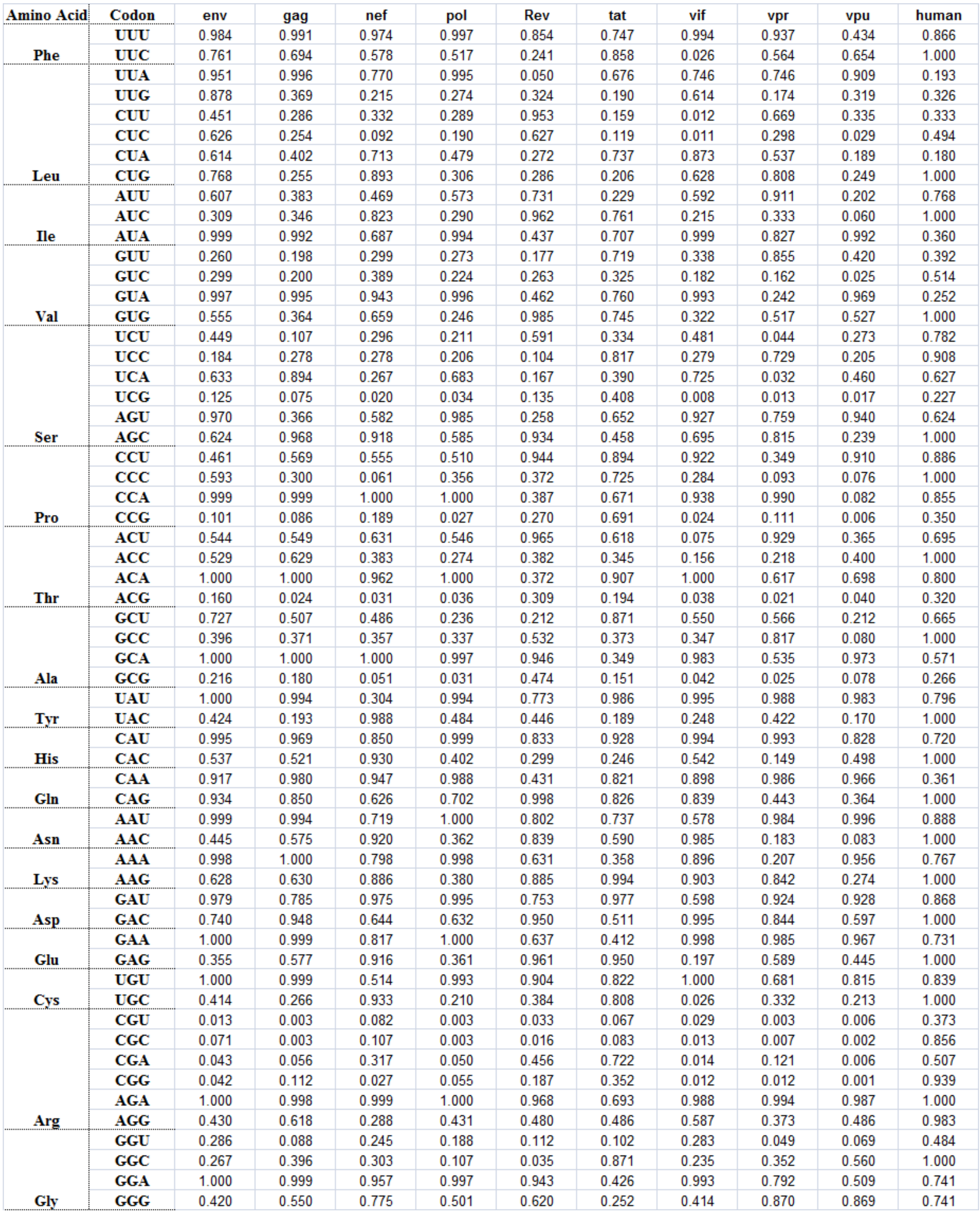
3.3 Results

3.3.1 Codon usage pattern of HIV-1 genes

The codon usage pattern of nine HIV-1 genes was normalized (refer Materials and Methods) by taking the total frequency for each of the 59 codons for 9 genes extracted from 241 whole genomes. Figure 3.1 shows the table of normalized codon usage data of 18 amino acids for 9 genes for all years, along with the host human. This data was analysed using PCA. Figure 3.2(a) shows the biplot of the first two components of PCA (accounting for 50.23% variance). The figure shows presence of a cline in the HIV-1 genes with respect to human. Of the four regulatory genes (*nef, rev, tat* and *vpr*), *rev* gene is closest to human followed by *tat*, *nef* and *vpr*, respectively. The three structural genes (*env, gag*, and *pol*) and the two other regulatory genes (*vif* and *vpu*) cluster away from human. This indicates that the codon usage patterns are different among the HIV-1 genes, and some of them are closer to the human host.

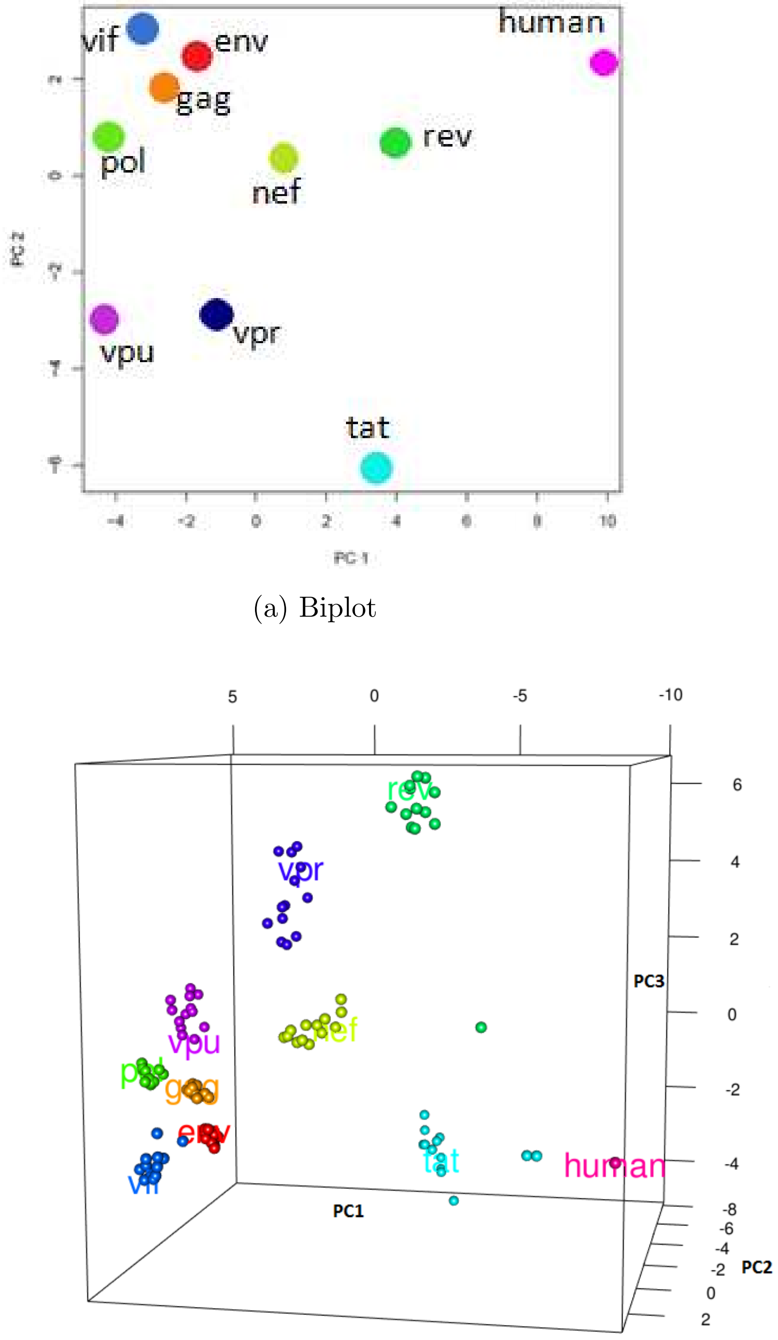
3.3.2 Temporal variability in the codon usage patterns of HIV-1 genes

In order to study if the non-random distribution of codon usage pattern of HIV-1 genes (as seen in Figure 3.2(a)) has any temporal trend, I performed a year-wise study of the normalized codon usage pattern of the HIV-1 genes from 1983 to 1995 using PCA. The first six Principal Components accounted for 89.65% of the variance in data. Figure 3.2(b) shows the first three PCs (accounting for 62.82% variance) for the data, where 13 points in PCA space (for 13 years) for each gene cluster close to each other. Four genes *env*, *gag*, *pol* and *vif* that clustered away from human in Figure 3.2(a), exhibit low cluster variance over the years - indicating comparatively stable (invariant) codon usage pattern. The *nef, rev, tat, vpr* and *vpu* genes exhibit high cluster variance indicating larger variation in their codon usage pattern over the 13 years. Since all gene sequences for each year were obtained from the same genomes, this variability in Figure 3.2(b) is representative of the intrinsic difference in codon usage, and not due to within-sample variations for any specific year. This analysis implies that the genes (*nef, rev, tat*, and *vpr*) of HIV-1 not only show differential segregation (with respect to human host), they also exhibit far more temporal variability in their codon usage patterns compared to the other genes.



Amino Acid	Codon	env	gag	nef	pol	Rev	tat	vif	vpr	vpu	human
Phe	UUU	0.984	0.991	0.974	0.997	0.854	0.747	0.994	0.937	0.434	0.866
	UUC	0.761	0.694	0.578	0.517	0.241	0.858	0.026	0.564	0.654	1.000
Leu	UUA	0.951	0.996	0.770	0.995	0.050	0.676	0.746	0.746	0.909	0.193
	UUG	0.878	0.369	0.215	0.274	0.324	0.190	0.614	0.174	0.319	0.326
	CUU	0.451	0.286	0.332	0.289	0.953	0.159	0.012	0.669	0.335	0.333
	CUC	0.626	0.254	0.092	0.190	0.627	0.119	0.011	0.298	0.029	0.494
	CUA	0.614	0.402	0.713	0.479	0.272	0.737	0.873	0.537	0.189	0.180
	CUG	0.768	0.255	0.893	0.306	0.286	0.206	0.628	0.808	0.249	1.000
Ile	AUU	0.607	0.383	0.469	0.573	0.731	0.229	0.592	0.911	0.202	0.768
	AUC	0.309	0.346	0.823	0.290	0.962	0.761	0.215	0.333	0.060	1.000
	AUA	0.999	0.992	0.687	0.994	0.437	0.707	0.999	0.827	0.992	0.360
Val	GUU	0.260	0.198	0.299	0.273	0.177	0.719	0.338	0.855	0.420	0.392
	GUC	0.299	0.200	0.389	0.224	0.263	0.325	0.182	0.162	0.025	0.514
	GUA	0.997	0.995	0.943	0.996	0.462	0.760	0.993	0.242	0.969	0.252
	GUG	0.555	0.364	0.659	0.246	0.985	0.745	0.322	0.517	0.527	1.000
Ser	UCU	0.449	0.107	0.296	0.211	0.591	0.334	0.481	0.044	0.273	0.782
	UCC	0.184	0.278	0.278	0.206	0.104	0.817	0.279	0.729	0.205	0.908
	UCA	0.633	0.894	0.267	0.683	0.167	0.390	0.725	0.032	0.460	0.627
	UCG	0.125	0.075	0.020	0.034	0.135	0.408	0.008	0.013	0.017	0.227
	AGU	0.970	0.366	0.582	0.985	0.258	0.652	0.927	0.759	0.940	0.624
	AGC	0.624	0.968	0.918	0.585	0.934	0.458	0.695	0.815	0.239	1.000
Pro	CCU	0.461	0.569	0.555	0.510	0.944	0.894	0.922	0.349	0.910	0.886
	CCC	0.593	0.300	0.061	0.356	0.372	0.725	0.284	0.093	0.076	1.000
	CCA	0.999	0.999	1.000	1.000	0.387	0.671	0.938	0.990	0.082	0.855
	CCG	0.101	0.086	0.189	0.027	0.270	0.691	0.024	0.111	0.006	0.350
Thr	ACU	0.544	0.549	0.631	0.546	0.965	0.618	0.075	0.929	0.365	0.695
	ACC	0.529	0.629	0.383	0.274	0.382	0.345	0.156	0.218	0.400	1.000
	ACA	1.000	1.000	0.962	1.000	0.372	0.907	1.000	0.617	0.698	0.800
	ACG	0.160	0.024	0.031	0.036	0.309	0.194	0.038	0.021	0.040	0.320
Ala	GCU	0.727	0.507	0.486	0.236	0.212	0.871	0.550	0.566	0.212	0.665
	GCC	0.396	0.371	0.357	0.337	0.532	0.373	0.347	0.817	0.080	1.000
	GCA	1.000	1.000	1.000	0.997	0.946	0.349	0.983	0.535	0.973	0.571
	GCG	0.216	0.180	0.051	0.031	0.474	0.151	0.042	0.025	0.078	0.266
Tyr	UAU	1.000	0.994	0.304	0.994	0.773	0.986	0.995	0.988	0.983	0.796
	UAC	0.424	0.193	0.988	0.484	0.446	0.189	0.248	0.422	0.170	1.000
His	CAU	0.995	0.969	0.850	0.999	0.833	0.928	0.994	0.993	0.828	0.720
	CAC	0.537	0.521	0.930	0.402	0.299	0.246	0.542	0.149	0.498	1.000
Gln	CAA	0.917	0.980	0.947	0.988	0.431	0.821	0.898	0.986	0.966	0.361
	CAG	0.934	0.850	0.626	0.702	0.998	0.826	0.839	0.443	0.364	1.000
Asn	AAU	0.999	0.994	0.719	1.000	0.802	0.737	0.578	0.984	0.996	0.888
	AAC	0.445	0.575	0.920	0.362	0.839	0.590	0.985	0.183	0.083	1.000
Lys	AAA	0.998	1.000	0.798	0.998	0.631	0.358	0.896	0.207	0.956	0.767
	AAG	0.628	0.630	0.886	0.380	0.885	0.994	0.903	0.842	0.274	1.000
Asp	GAU	0.979	0.785	0.975	0.995	0.753	0.977	0.598	0.924	0.928	0.868
	GAC	0.740	0.948	0.644	0.632	0.950	0.511	0.995	0.844	0.597	1.000
Glu	GAA	1.000	0.999	0.817	1.000	0.637	0.412	0.998	0.985	0.967	0.731
	GAG	0.355	0.577	0.916	0.361	0.961	0.950	0.197	0.589	0.445	1.000
Cys	UGU	1.000	0.999	0.514	0.993	0.904	0.822	1.000	0.681	0.815	0.839
	UGC	0.414	0.266	0.933	0.210	0.384	0.808	0.026	0.332	0.213	1.000
Arg	CGU	0.013	0.003	0.082	0.003	0.033	0.067	0.029	0.003	0.006	0.373
	CGC	0.071	0.003	0.107	0.003	0.016	0.083	0.013	0.007	0.002	0.856
	CGA	0.043	0.056	0.317	0.050	0.456	0.722	0.014	0.121	0.006	0.507
	CGG	0.042	0.112	0.027	0.055	0.187	0.352	0.012	0.012	0.001	0.939
	AGA	1.000	0.998	0.999	1.000	0.968	0.693	0.988	0.994	0.987	1.000
	AGG	0.430	0.618	0.288	0.431	0.480	0.486	0.587	0.373	0.486	0.983
Gly	GGU	0.286	0.088	0.245	0.188	0.112	0.102	0.283	0.049	0.069	0.484
	GGC	0.267	0.396	0.303	0.107	0.035	0.871	0.235	0.352	0.560	1.000
	GGA	1.000	0.999	0.957	0.997	0.943	0.426	0.993	0.792	0.509	0.741
	GGG	0.420	0.550	0.775	0.501	0.620	0.252	0.414	0.870	0.869	0.741

Figure 3.1: Normalized codon usage data of 18 amino acids for nine HIV-1 genes from 241 whole genomes for 13 years and average human codon usage



(a) Biplot

(b) Plot of 3 principal components

Figure 3.2: Multivariate Analysis of the codon usage patterns of nine HIV-1 genes and human. (a) Biplot of the first two Principal Components for nine HIV-1 genes and human. (b) First three Principal Components plot for human genes and nine HIV-1 genes for 13 years. Different colours are assigned for different HIV-1 genes and human. Each gene-specific cluster contains 13 data points corresponding to the genomes from 13 years.

Chapter 4

Hierarchical Model for Genome Wide Association Study (GWAS) of *Psoriasis*

4.1 Bayesian Statistics

Scientific hypotheses typically are expressed through probability distributions for observable scientific data. These probability distributions depend on unknown quantities called parameters. In the Bayesian paradigm, current knowledge about the model parameters is expressed by placing a probability distribution on the parameters, called the 'prior distribution', often written as

$$p(\theta)$$

When new data become available, the information they contain regarding the model parameters is expressed in the 'likelihood', which is proportional to the distribution of the observed data given the model parameters, written as

$$p(y|\theta)$$

This information is then combined with the prior to produce an updated probability distribution called the 'posterior distribution', on which all Bayesian inference is based. Bayes' Theorem, an elementary identity in probability theory, states how the update is done mathematically: the posterior is proportional to the prior times the likelihood, or more precisely,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} \tag{4.1}$$

where the denominator is the 'marginal likelihood'.

The fullest version of the Bayesian paradigm casts statistical problems in the framework of decision making. It entails formulating subjective prior probabilities to express pre-existing information, careful modelling of the data structure, checking and allowing for uncertainty in model assumptions, formulating a set of possible decisions and a utility function to express how the value of each alternative decision is affected by the unknown model parameters (http://bayesian.org/Bayes)

4.2 Some Genetics Concepts

4.2.1 Single Nucleotide Polymorphisms (SNP)

Single Nucleotide Polymorphisms, frequently called SNPs, are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. SNPs occur normally throughout a persons DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the genes function.

Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individuals response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer.

4.2.2 Genotype and Phenotype

The genotype is a person's complete heritable genetic identity. Personal genome sequencing reveals the unique genome of each individual. However, the word genotype can also refer just to a particular gene or set of genes carried by an individual. For example, if a person carries a mutation that is linked to diabetes, then that may just be referred to its genotype with respect to this mutation without consideration of all the other gene variants that it may carry.

In contrast, the phenotype is a description of the person's actual physical characteristics. This includes straight-forward visible characteristics like, height and eye color, but can also include overall health, disease history, and even behavior and general disposition. However, not all phenotypes are a direct result of a person's genotype. Most phenotypes are influenced by both the genotype and by the unique circumstances in which a person has lived its life. These two inputs are often referred to as nature, the unique genome one carries, and nurture, the environment in which one has lived one's life.

4.2.3 Allele

An allele is a variant form of a gene. Some genes have a variety of different forms, which are located at the same position, or genetic locus, on a chromosome. Humans are called diploid organisms because they have two alleles at each genetic locus, with one allele inherited from each parent. Each pair of alleles represents the genotype of a specific gene. Genotypes are described as homozygous if there are two identical alleles at a particular locus and as heterozygous if the two alleles differ. Alleles contribute to the organism's phenotype, which is the outward appearance of the organism.

Some alleles are dominant or recessive. When an organism is heterozygous at a specific locus and carries one dominant and one recessive allele, the organism will express the dominant phenotype. Alleles can also refer to minor DNA sequence variations between alleles that do not necessarily influence the gene's phenotype.

4.2.4 Linkage Disequilibrium (LD)

Linkage disequilibrium (LD) is a property of SNPs on a contiguous stretch of genomic sequence that describes the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP within a population. The term LD was coined by population geneticists in an attempt to mathematically describe changes in genetic variation within a population over time. It is related to the concept of chromosomal linkage, where two markers on a chromosome remain physically joined through generations of a family. Recombination events within a family from generation to generation break apart chromosomal segments. This effect is amplified through generations, and in a population of fixed size undergoing random mating, repeated random recombination events will break apart segments of contiguous chromosome (containing linked alleles) until eventually all alleles in the population are in linkage equilibrium or are independent. Thus, linkage between markers on a population scale is referred to as linkage disequilibrium.

4.3 Genome-Wide Association Studies (GWAS)

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

4.3.1 Dataset for Case-Control Studies

Let, there be p SNPs considered in GWAS data for n individuals (n_1 Controls and n_2 Cases such that $n = n_1 + n_2$).

Genotype Data - Let, $X_{n \times p}$ be the Genotype data matrix where X_{ij} be the genotype of i^{th} individual for j^{th} SNP. The three possible genotypes at a particular locus are encoded in the following way:
AA - 0, Aa - 1, and aa - 2
Where 'A' is a major allele and 'a' is a minor allele.

Phenotype Data - Let, Y_i be the phenotype of the i^{th} individual. In case-control studies it is given as:
$Y_i = 0$ for control and $Y_i = 1$ for case
We applied the hierarchical model on GWAS data of *Psoriasis* disease which comprises of 586266 SNPs for 688 Controls and 926 Cases [$www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id = phs000019.v1.p1$].

4.3.2 Quality Control Check

The following quality control checks were done on GWAS genotype data :

1. Dropped the SNPs having $\geq$ 90% individuls data and also the individuals having $\geq$ 90% missing SNP data.

2. Dropped the SNPs having Minor Allele Frequency (MAF) $\leq$ 1%

3. Replaced the missing values for each SNPs by the average genotype value.

4. Dropped the SNPs which are monomorphic or for which variance equals to zero.

The above quality control check is done on both the control genotypic data and also in cases genotypic data. Then only those SNPs are included in the analysis which are left in both the datasets after quality control check.

4.3.3 PCA on GWAS data

Let, after the quality control check, we are left with $X_{1(t \times n_1)}$ (Control genotype matrix) and $X_{2(t \times n_2)}$ (Case genotype matrix). Since some SNPs can be correlated as an effect of LD, PCA is performed on Genotypic data of controls (X_1^T), which will be used to solve population stratification problem later. Let, $V_{t \times n_1}$ be the loading matrix of this PCA. Then, the two Score (S) matrices are computed as

$$S_{controls} = X_1^T V \qquad\qquad S_{cases} = X_2^T V \tag{4.2}$$

If the proportion of cases in the population is very low, only loading matrix of control-PCA is used. There are a total n_1 PCs. After concatenating the two matrices we get a Score for $S_{(n_1+n_2) \times n_1}$ for n_1 controls and n_2 cases.
Let, $Y_{(n_1+n_2) \times 1}$ be the disease status data matrix, where the first n_1 entries are $0's$ representing *controls*, and the remaining n_2 entries are $1's$ representing *cases*. Since the data is now binary, Logistic regression is performed on Y and S by taking first l components from S, where l varies from 1 to n_1.

4.4 Hierarchical Model

In association studies, statistical models have been developed to identify shared characteristics of SNPs that influence a disease. In a hierarchical modelling framework, the probability that a given SNP is a causal can then depend on these characteristics. In the GWAS context, the number of loci unambiguously associated with a given trait or a disease has historically been very small. The goal of this approach is to identify the shared characteristics of SNPs and find those SNPs that causally influence the trait or disease [Pic14].

4.4.1 Method

The basic building block of the modelling process is the linear regression model. Let, $\bar{y}$ be the vector of phenotypes and $\bar{g}$ be the vector of genotypes. A standard additive linear model is -

$$E(y_i) = \alpha + \beta g_i \tag{4.3}$$

The Null and Alternative hypotheses are defined as

$$\begin{aligned} H_0 &: \ \beta = 0 \\ H_1 &: \ \beta \neq 0 \end{aligned}$$

Let us consider a set of M SNPs, each of which has been genotyped in N individuals in a GWAS. Because of LD, there will be many SNPs whose associations are not causal. However, these will all be restricted to a block around the truly causal site. Hence, the genome is split into contiguous blocks of size K SNPs such that there are M/K blocks. The block size is chosen to be much larger than the extent of LD in the population.

Let, Π_k be the prior probability that block k contains a causal SNP associated with the trait. Let Δ_k be the parameter for the region k denoting if the region contains a causal SNP or not (having values 1 or 0). The probability of the data (the set of observed phenotypes) is then given by

$$P(\bar{y}) = \prod_{k=1}^{M/K} (1 - \Pi_k) P_k^0 + \Pi_k P_k^1 \tag{4.4}$$

where

$$\begin{aligned} P_k^0 &= P(Data|\Delta_k = 0) \\ P_k^1 &= P(Data|\Delta_k = 1) \\ P_k^1 &= \sum_{i \in S_k} \pi_{ik} P_{ik}^1 \end{aligned}$$

where S_k is the set of SNPs in the block k. Assume that there is a single association in block k. Let δ_{ik} be the parameter for the i^{th} SNP in the k^{th} region to be causal or not (1/0). So,

$$\pi_{ik} = P(\delta_{ik} = 1 | \Delta_k = 1)$$
$$P_{ik}^1 = P(Data | \delta_{ik} = 1)$$

4.4.2 Computing Bayes factor

Let, $\hat{\beta}$ be the maximum-likelihood estimate of β such that

$$\hat{\beta} \sim N(\beta, V)$$
$$\beta \sim N(0, W)$$

Then, Bayes factor (BF) is given as

$$BF = \frac{P(\bar{y}|\bar{g}, H_1)}{P(\bar{y}|\bar{g}, H_0)} = \frac{\int P(\hat{\beta}|\beta)\Pi(\beta)d\beta}{P(\hat{\beta}|\beta = 0)} \tag{4.5}$$

where

$$P(\hat{\beta}|\beta) = \frac{1}{\sqrt{2V^2\pi}} \exp^{-\frac{(\hat{\beta}-\beta)^2}{2V^2}}$$
$$\Pi(\beta) = \frac{1}{\sqrt{2W^2\pi}} \exp^{-\frac{(\beta)^2}{2W^2}}$$
$$P(\hat{\beta}|\beta = 0) = \frac{1}{\sqrt{2V^2\pi}} \exp^{-\frac{(\hat{\beta})^2}{2V^2}}$$

The second equality in equation 4.5 is computed in Wakefield [Wak09], which uses summary of the linear regression for finding the Bayes factor.

4.4.3 Fitting the Model

When the terms above are combined, the likelihood of the data can be written as,

$$L(\bar{y}|\theta) = \prod_{k=1}^{M/K} (1 - \Pi_k)P_k^0 + \Pi_k \sum_{i=1}^{K} \pi_{ij} P_{ik}^1 \tag{4.6}$$
$$L(\bar{y}|\theta) = \prod_{k=1}^{M/K} P_k^0[(1 - \Pi_k)P_k^0 + \Pi_k \sum_{i=1}^{K} \pi_{ij} BF_i] \tag{4.7}$$

where θ contains all the parameters of the model. This function is maximised to get $\hat{\Pi}_k$.

4.4.4 Posterior Probabilities of Association (PPA)

Once the model has been fit, it produces empirical estimates of the prior probability that region k contains an association, $\hat{\Pi}_k$, and of the prior probability that SNP i is the causal one, $\hat{\pi}_{ik}$ (on the condition that there is an association). Let us define a Bayes factor that summarizes the evidence of association in the region as,

$$BF_k^R = \sum_{i \in S_k} \hat{\pi}_{ik} BF_i = P_k^1 / P_k^0$$

where

$$BF_i = \frac{P_{ik}^1}{P_k^0}$$

Now the posterior probability that region k contains an association is given by

$$\begin{aligned} P(\Delta_k = 1 | Data) &= \frac{P(Data|\Delta_k = 1) P(\Delta_k = 1)}{P(Data)} \\ &= \frac{P_k^1 \Pi_k}{(1 - \Pi_k) P_k^0 + \Pi_k P_k^1} \\ &= \frac{\hat{\Pi}_k BF_k^R / (1 - \hat{\Pi}_k)}{1 + \hat{\Pi}_k BF_k^R / (1 - \hat{\Pi}_k)} \end{aligned}$$

We can also define the posterior probability that any given SNP i in region k is the causal one under the model

$$\begin{aligned} P(\delta_{ik} = 1 | Data, \Delta_k = 1) &= \frac{P(Data|\delta_{ik} = 1, \Delta_k = 1) P(\delta_{ik} = 1 | \Delta_k = 1)}{P(Data|\Delta_k = 1)} \\ &= \frac{P_{ik}^1 \hat{\pi}_{ik}}{P_k^1} \\ &= \frac{P_{ik}^1 \hat{\pi}_{ik}}{\sum_{j \in S_k} \hat{\pi}_{jk} P_{jk}^1} \\ &= \frac{P_{ik}^1 \hat{\pi}_{ik} / P_k^0}{\sum_{j \in S_k} \hat{\pi}_{jk} P_{jk}^1 / P_k^0} \\ &= \frac{\hat{\pi}_{ik} BF_i}{\sum_{j \in S_k} \hat{\pi}_{jk} BF_j} \end{aligned}$$

Hence, the Posterior probability that any given SNP is causal is given by

$$\boxed{P(\delta_{ik} = 1 | Data) = P(\Delta_k = 1 | Data) P(\delta_{ik} = 1 | Data, \Delta_k = 1)} \tag{4.8}$$

4.4.5 Population Stratification (PS)

Population stratification [ZZQ+15] refers to the presence of a systematic difference in allele frequencies between subpopulations in a study due to ancestry difference between the study subjects.

Problems caused due to PS in GWAS-

1. Unrecognized population stratification can lead to both false-positive and false-negative findings.

2. It can obscure the true association signals if not appropriately corrected.

Recognizing the issue of population stratification induced by population structures, various methods have been developed to control for population stratification. Two approaches are

1. Genomic control [DR99] and

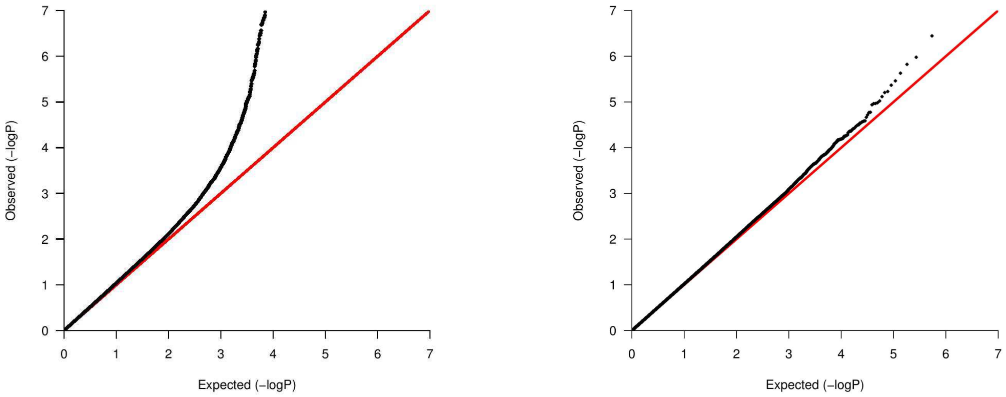
2. EIGENSTRAT [PPR06]

Genomic control -

The genomic control method corrects for stratification by adjusting association statistics with an overall inflation factor obtained from a set of random markers that are not associated with the phenotypes of interest. However, some markers differ in their allele frequencies across ancestral populations more than others. Thus, the uniform adjustment may be insufficient at markers having strong differentiation across ancestral populations and may be superfluous at markers lacking such differentiation.

EIGENSTRAT -

The current state-of-the-art approach for the correction of population stratification is EIGENSTRAT, which computes principal components for SNPs across the genome to identify population structure. In this approach, a small number of top principal components will capture the main axes of genetic variation in the study subjects. Correction for population stratification is carried out by including these top principal components as covariates in a regression framework. In this study, I test each SNPs association with the outcome of interest by building a logistic regression model that includes the specific SNP as one factor and the selected PCs as covariates.



(a) Q-Q plot before PS correction

(b) Q-Q plot after PS correction

Figure 4.1: Q-Q plots before and after PS correction. (a) P-values from logistic regression on *Psoriasis* data. (b) P-values obtained after correcting population stratification problem by using 20 PCs in EIGENSTRAT correction (equation (4.9))

4.4.6 Association testing using Logistic Regression models

To perform the association analysis for each SNP, a logistic regression model was used with the specific SNP as one factor and the PCs from the robust method as the covariates for the *Psoriasis* data.

$$logit(Y) = \alpha + \beta g + \gamma X \tag{4.9}$$

where Y represents the binary response variable (such as the disease status - 0/1), g represents the genotype value of the specific SNP, X represents the PCs from the EIGENSTRAT method. The plots show the results obtained for the specific data (see Section 4.3.1) analysed.

4.4.7 Quantile-Quantile (Q-Q) plot

QQ plots are used to visualize the relationship between the expected and observed distributions of SNP-level test statistics. In Figure 4.1(a), observed $-\log(p-value)$ is plotted against expected $-\log(p-value)$ by using p-values from the logistic regression from equation (4.3). The tail of the distribution of observed values deviate considerably from expected. The tail is brought closer to the $y = x$ line after accounting for potential confounding by population stratification (as shown in 4.1(b)) in which p-values are obtained by equation (4.9) . Even after that, the two lines do not overlap, which implies that some form of association is present in the data.

Genomic Inflation Factor

The genomic inflation factor (λ) is defined as the ratio of the median of the empirically observed distribution of the test statistic (χ^2) to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate.

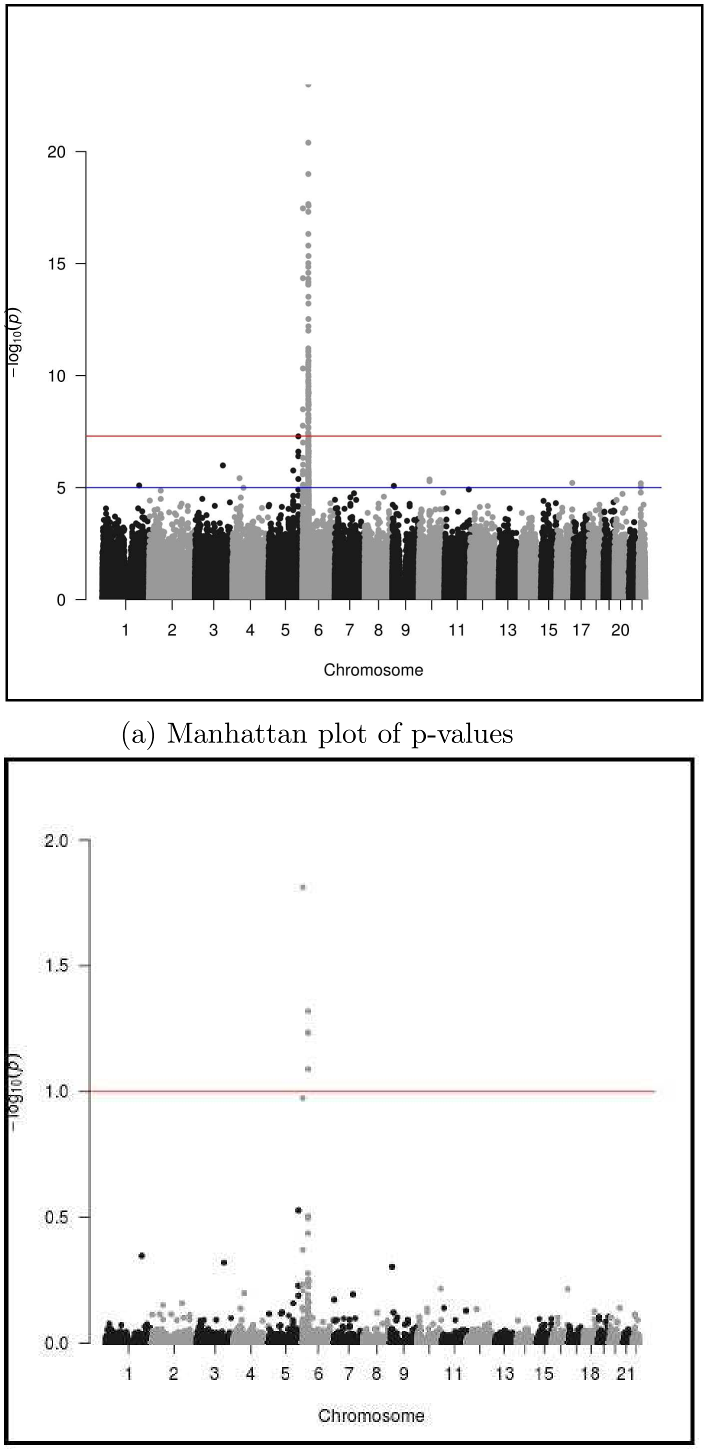
$$\lambda = median(\chi^2)/0.456$$

$$\chi^2_{adjusted} = \chi^2/\lambda$$

Inflated λ values or residual deviations in the QQ plot may point to undetected sample duplications, unknown familial relationships, a poorly calibrated test statistic, systematic technical bias or gross population stratification. For *Psoriasis* data, I computed genomic inflation factor as 1.87, which shows the presence of population stratification. So, we corrected it by using EIGENSTRAT approach with first 3 principal components (equation (4.9)).

4.4.8 Manhattan plots

Manhattan plots are used to visualize GWA significance level by chromosome location as shown in Figure 4.2. Here, each dot corresponds to a single SNP. The x-axis represents gene coordinates, and the numbers shown correspond to chromosome numbers. The y-axis is the negative of the log(p-value) in Figure 4.2(a) and the negative of the log(1-PPA) in Figure 4.2(b), where PPA stands for Posterior Probability of Association for SNP. The red line in Figure 4.2(a) indicates the Bonferonni corrected significance threshold $-\log(5 \times 10^{-8})$. The black line is a less stringent suggestive association threshold $-\log(5 \times 10^{-6})$ that we use as an indicator of a suggestive association and requiring further validation.



(a) Manhattan plot of p-values

(b) Manhattan plot of (1-PPA)

Figure 4.2: Manhattan plots of p-values and (1-PPA) for the SNPs in the *Psoriasis* data-set. (a) By using p-values from frequentist approach which shows marginal association using logistic regression (equation (4.9)). (b) By using (1-PPA) from Empirical Bayes approach (Hierarchical model). The red line is for PPA = 0.9

.

4.5 Summary

▶ We replicated the approach of Pickrell 2014 [Pic14], implemented it in R and applied it on *Psoriasis* GWAS data.

▶ SNPs rs2844627, rs13191258, rs12191877, rs9267673 on the 6^{th} chromosome were found to have significant posterior probability of being causal (PPA$\geq$ 0.9).

▶ Due to high correlation of SNPs in local neighbourhoods, locating causal SNPs is a hard problem in GWAS. The strategy of assuming one causal SNP per region (prior) helps significantly to solve this problem.

Bibliography

[A08] Wagner A, *Neutralism and selectionism: A network-based reconciliation*, Nat Rev Genet **9** (2008), 965974.

[AC91] SJ Arrigo and IS Chen, *Rev is necessary for translation but not cytoplasmic accumulation of HIV-1 vif, vpu, and env/vpu 2 rnas*, Genes Dev **5** (1991), 808819.

[AH04] JL Anderson and TJ Hope, *HIV accessory proteins and surviving the host cell.*, Curr HIV/AIDS Rep **1** (2004), 4753.

[CH09] JV Chamary and LD Hurst, *How trivial DNA changes can hurt health*, Sci Am **30** (2009), 4653.

[CPH06] JV Chamary, JL Parmley, and LD Hurst, *Hearing silence: non-neutral evolution at synonymous sites in mammals*, Nat Rev Genet **7** (2006), 98108.

[DR99] B Devlin and K Roeder, *Genomic control for association studies.*, Biometrics **55(4)** (1999), 997–1004.

[DSR+86] Andrew I Dayton, Joseph G Sodroski, Craig A Rosen, Wei Chun Goh, and William A Haseltine, *The trans-activator gene of the human T cell lymphotropic virus type iii is required for replication*, Cell **44** (1986), no. 6, 941–947.

[Efr05] Bradley Efron, *Local false discovery rates.*

[EP04] Rocha EP, *Codon usage bias from tRNAs point of view: Redundancy, specialization, and efficient decoding for translation optimization.*, Genome Res **14** (2004), 22792286.

[FFJ+86] Amanda G Fisher, Mark B Feinberg, Steven F Josephs, Mary E Harper, Lisa M Marselle, Gregory Reyes, Matthew A Gonda, Anna Aldovini, Christine Debouk, Robert C Gallo, et al., *The trans-activator gene of HTLV-III is essential for virus replication*, Nature **320** (1986), no. 6060, 367–371.

[Hot33] H Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology **24(6)** (1933), 417–441, http://dx.doi.org/10.1037/h0071325.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *TheElements of Statistical Learning*, Springer, 2009.

[Ike85] T Ikemura, *Codon usage and tRNA content in unicellular and multicellular organisms*, Mol Biol Evol **2** (1985), 13–34.

[JH03] G Jenkins and E Holmes, *The extent of codon usage bias in human RNA viruses and its evolutionary origin*, Virus Res **92** (2003), 1–7.

[KCT$^+$03] GH Kijak, JR Currier, S Tovanabutra, JH Cox, NL Michael, SA Wegner, DL Birx, and FE McCutchan, *Lost in translation: implications of HIV-1 codon usage for immune escape and drug resistance.*, AIDS reviews **6** (2003), no. 1, 54–60.

[LNKP08] Julius B Lucks, David R Nelson, Grzegorz R Kudla, and Joshua B Plotkin, *Genome landscapes and bacteriophage codon usage*, PLoS Comput Biol **4** (2008), no. 2, e1000001.

[McL04] G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*, Wiley, 2004, ISBN 0-471-69115-1. MR 1190469.

[MS97] RH Miller and N Sarver, *HIV accessory proteins as therapeutic targets*, Nature Med **3** (1997), 389394.

[NGI00] Y Nakamura, T Gojobori, and T Ikemura, *Codon usage tabulated from international DNA sequence databases: status for the year 2000*, Nuc Acids Res **28** (2000), 292.

[Pic14] JK Pickrell, *Joint analysis of functional genomic data and Genome-wide Association Studies of 18 Human Traits*, American Journal of Human Genetics **94(4)** (2014), 559–573.

[PPR06] N Patterson, AL Price, and D Reich, *Population Structure and Eigenanalysis*, PLoS Genet **2(12)** (2006), e190.

[PRL04] JB Plotkin, H Robins, and AJ Levine, *Tissue-specific codon usage and the expression of human genes*, Proc Natl Acad Sci USA **101** (2004), 1258812591, DOI: 10.1073/pnas.0404957101.

[PT02] G Perriere and J Thioulouse, *Use and misuse of correspondence analysis in codon usage studies*, Nuc Acids Res **30** (2002), 4548–4555.

[SST05] H Suzuki, R Saito, and M Tomita, *A problem in multivariate analysis of codon usage data and a possible solution.*, FEBS Lett **579** (2005), 6499–6504.

[Sto08] N Stoletzki, *Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures*, BMC Evol Biol **8** (2008), 224.

[Tro95] D Trono, *HIV accessory proteins: leading roles for the supporting cast*, Cell **82** (1995), 189192.

[UH03] AO Urrutia and LD Hurst, *The signature of selection mediated by expression on human genes*, Genome Res **13** (2003), 22602264.

[Wak09] J Wakefield, *Bayes factors for gnome-wide association studies*, Genet. Epidemiol. **33** (2009), 79–86.

[Wil03] S Williamson, *Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression*, Mol Biol Evol **20** (2003), 13181325.

[WPB$^+$05] Scott Williamson, Steven M Perry, Carlos D Bustamante, Maria E Orive, Miles N Stearns, and John K Kelly, *A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients*, Molecular biology and evolution **22** (2005), no. 3, 456–468.

[ZB86] Jerrold H Zar and TJ Breen, *Contemporary textbooks on Multlvariate Statlstlcal Analysis: A panoramic appraisal and critique*, 1986.

[ZZQ$^+$15] Ping Zeng, Yang Zhao, Cheng Qian, Liwei Zhang, Ruyang Zhang, Jianwei Gou, Jin Liu, Liya Liu, and Feng Chen, *Statistical analysis for genome-wide association study*, Journal of biomedical research **29** (2015), no. 4, 285.

APPENDIX : Programs

1. Comparison of LASSO and FDR methods

```
#Simulations:
set.seed(150)
n <- 500
p <- 2000
pi <- seq(0.01,0.15,length.out=5)
t <- floor(p*pi)
sigma <- matrix(0,p,p)
diag(sigma) <- 1
library(mvtnorm)
xs <-  rmvnorm(n, mean=rep(0,p),sigma=sigma)
sig.true <- 20
#LASSO:
lasso_mat <- function(y)
{
library(glmnet)
fit <- glmnet(x=xs,y=y)
seq <- unique(fit$df)
lambda.new <- NULL
for(r in 1:length(seq))
{
lambda.new[r] <- fit$lambda[which(fit$df==seq[r])[1]]
}
fit1 <- glmnet(x=xs,y=y,lambda=c(lambda.new,10^(seq(log10(0.7),
    log10(0.002), length.out=50)),0))
mat1 <- as.matrix(coef(fit1))
seq1 <- unique(fit1$df)
mat1 <- mat1[-1,]
mat_lasso <- t(mat1)
for(k in 1:ncol(mat_lasso)){mat_lasso[mat_lasso[,k] != 0,k] <- 1}
list(mat=mat_lasso,df=seq1)
}
#FDR:
fdr_mat <- function(y,vec)
{
```

```
library(locfdr)
bet <- NULL
for(k in 1:p)
{
bet[k] <- summary(lm(y~xs[,k]))$coef[2,3]
}
no<- as.list(rep(NA,length(vec)))
sort_bet <- sort(abs(bet), decreasing=TRUE)
for(q in 1:length(vec)){no[[q]]<- which(abs(bet) >= sort_bet[vec[q
    ]])}
mat_fdr <- matrix(0,length(vec),p)
for(l in 1:length(vec))
{
  for(m in 1:length(no[[l]])){mat_fdr[l,no[[l]][m]] <- 1}
}
return(mat_fdr)
}
summarize <- function(type)
{
TP <- TN <- FP <- FN <- nvar <- as.list(rep(NA, length(t)))
sens <- spec <- TPR <- FPR <- as.list(rep(NA, length(t)))
len <- rep(NA, length(t))
for(j in 1:length(t))
{ tt <- t[j]
  mat_type <- if(type=="lasso") arr.list[[1]][[j]] else arr.list
      [[2]][[j]]
  len[j] <- nrow(mat_type)
  gamma <- t(as.matrix(c(rep(1, tt), rep(0, p - tt))))
  sens[[j]] <- spec[[j]] <- FPR[[j]] <- TPR[[j]] <- numeric(len[j])
  for(i in 1:len[j])
  { TP <- sum(mat_type[i,1:tt]==1)
    TN <- sum(mat_type[i,(tt+1):p]==0)
    FP <- sum(mat_type[i,(tt+1):p]==1)
    FN <- sum(mat_type[i,1:tt]==0)
     print(TP/FP)
     sens[[j]][i] <- TP/ (TP + FN)
     spec[[j]][i] <- TN / (TN + FP)
    FPR[[j]][i] <- 1 - spec[[j]][i]
    TPR[[j]][i] <- sens[[j]][i]
   }
}
list(Spec=spec, Sens=sens, FPR=FPR, TPR=TPR, len=len)
}
#plots
plots <- function(type, curve_name, res)
```

```
{
for(j in 1:length(t))
{
  tt <- t[j]
  first <- (type == "lasso" && j == 1)
  x_var <- y_var <- numeric(res$len[j])
  for(i in 1:res$len[j])
  { if(curve_name=="sens")
    {
    x_var[i] <- vect[[j]][i]
    y_var[i] <- res$Sens[[j]][i]
    x_lab <- "df"
    y_lab <- "sensitivity"
    x_lim <- c(0,p)
    plot_name <- "Sensitivity vs #Variables selected"}
    else if(curve_name=="spec")
    {x_var[i] <- vect[[j]][i]
    y_var[i] <- res$Spec[[j]][i]
    x_lab <- "df"
    y_lab <- "specificity"
    x_lim <- c(0,p)
    plot_name <- "Specificity vs #Variables selected"}
    else if(curve_name=="ROC")
    {x_var[i] <- res$FPR[[j]][i]
    y_var[i] <- res$TPR[[j]][i]
    x_lab <- "FPR"
    y_lab <- "TPR"
    x_lim <- c(0,1)
    plot_name <- "ROC Curve"}
  }
  if(first) plot(x_var, y_var, "l",xlab=x_lab,ylab=y_lab,xlim=x_lim
      ,ylim=c(0,1),main=plot_name, lty=1)
  if(!first && type=="lasso")lines(x_var,y_var,"l",xlab=x_lab,ylab=
      y_lab,main=plot_name,lty=j)
  if(!first && type=="fdr")lines(x_var,y_var,"l",xlab=x_lab,ylab=y_
      lab,main=plot_name,col="RED",lty=j)
}
}
if(TRUE)
{
y <- matrix(NA, length(t), n)
for(j in 1:length(t))
{
  beta <- rnorm(t[j], mean=1, sd=0.5)
  beta.true <- c(beta, rep(0, (p - t[j])))
  plot(density(beta.true))
  y[j, ] <- rnorm(n, mean=drop(xs $\%$*$\%$ beta.true), sd=sig.true
     )
}
vect <- as.list(rep(NA, length(t)))
arr.list <- as.list(c(NA, NA))
arr.list[[1]] <- as.list(1:length(t))
arr.list[[2]] <- as.list(1:length(t))
for(j in 1:length(t))
{
  lasso.res <- lasso_mat(y[j, ])
  vect[[j]] <- lasso.res$df
  fdr.res <- fdr_mat(y[j, ], vec=vect[[j]])
  arr.list[[1]][[j]] <- lasso.res$mat
  arr.list[[2]][[j]] <- fdr.res
}
}
res1 <- summarize("lasso")
res2 <- summarize("fdr")
pdf("lasso+fdr_wrt_t_univ(fdr)_beta.pdf")
plots("lasso","sens", res=res1)
plots("fdr","sens", res=res2)
plots("lasso","spec", res=res1)
plots("fdr","spec", res=res2)
plots("lasso","ROC", res=res1)
plots("fdr","ROC", res=res2)
dev.off()
```

2. Three PC Plot for codon usage data

```
library(xlsx)
library(rgl)
data1 <- read.xlsx("yearwise.xlsx",sheetIndex=1) #input the codon
   usage data
data1 <- data1[1:59,3:12]
data1 <- t(data1)
data1 <- as.matrix(data1)
j <- 2
while(j <= 13)
{
data <- read.xlsx("yearwise.xlsx",sheetIndex=j)
data <- data[1:59,3:12]
data <- t(data)
data <- as.matrix(data)
data1 <- rbind(data1,data)
j <- j+1
}
colour <- c(rainbow(10),rainbow(10),rainbow(10),rainbow(10),rainbow
    (10),rainbow(10),rainbow(10),rainbow(10),rainbow(10),rainbow(10)
    ,rainbow(10),rainbow(10),rainbow(10))
pc <- prcomp(data1,center=TRUE,scale=TRUE)
print(pc)
print(pc$x)
plot3d(pc$x[,1],pc$x[,2],pc$x[,3],col=colour,type=”s”, size=0.7)
texts3d(pc$x[1:10,1],pc$x[1:10,2],pc$x[1:10,3],texts=c(”env”,”gag”,
    ”nef”,”pol”,”rev”,”tat”,”vif”,”vpr”,”vpu”,”human”),col=rainbow
    (10), cex=2)
```