Comparing Feature Selection Algorithms Using Microarray Data

by

Timothy Tao Hin Law

B.Sc., University of British Columbia, 2003

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Comparing Feature Selection Algorithms Using Microarray Data

by

Timothy Tao Hin Law

B.Sc., University of British Columbia, 2003

Supervisory Committee

Dr. M. Lesperance, Supervisor
(Department of Mathematics and Statistics)

Dr. M. Tsao, Departmental Member
(Department of Mathematics and Statistics)

Dr. J. Zhou, Departmental Member
(Department of Mathematics and Statistics)

Abstract

Supervisory Committee

Dr. M. Lesperance, Supervisor
(Department of Mathematics and Statistics)

Dr. M. Tsao, Departmental Member
(Department of Mathematics and Statistics)

Dr. J. Zhou, Departmental Member
(Department of Mathematics and Statistics)

In this thesis study, three different feature selection methods, LASSO, SLR, and SMLR, were tested and compared using microarray fold change data. Two real datasets were used to first investigate and compare the ability of the algorithms in selecting feature genes on data under two conditions. It was found that SMLR was quite sensitive to its parameter, and was more accurate in selecting differentially expressed genes when compared to SLR and LASSO. In addition, the model coefficients generated by SMLR had a close relationship with the magnitude of fold changes. Also, SMLR's ability in selecting differentially expressed genes with data that had more than two conditions was shown to be successful. The results from simulation experiments agreed with the results from the real dataset experiments. Additionally, it was found that different proportions of differentially expressed genes in the data did not affect the performance of LASSO and SLR, but the number of genes selected by SMLR increased with the proportion of regulated genes. Also, as the number of replicates used to build the model increased, the number of genes selected by SMLR increased. This applied to both correctly and incorrectly selected genes. Furthermore, it was found that SMLR performed the best in identifying future treatment samples.

Table of Contents

Supervisory Committee **ii**

Abstract **iii**

Table of Contents **iv**

List of Tables **vi**

List of Figures **ix**

Acknowlegements **xii**

1 Introduction **1**

1.1 Microarray 1

2 Literature Review **5**

2.1 Least Absolute Shrinkage and Selection Operator (LASSO) 5

2.2 Sparse Logistic Regression (SLR) 6

2.3 Sparse Multinomial Logistic Regression (SMLR) 10

2.4 Model Setup (LASSO, SLR, and SMLR) 14

3 Intersex Data Experiment **16**

3.1 Male/Female Experiment 16

3.2 Intersex/Female Experiment 26

3.3 Intersex/Male Experiment 36

3.4 Male/Female/Intersex Experiment 46

4 Breast Cancer Data Experiment **48**

4.1 ER+/ER- Experiment 49

4.2 LN-/LN+ Experiment 57

4.3 ER+LN+/ER+LN-/ER-LN+/ER-LN- Experiment 65

5 Simulated Data Experiment **68**

5.1 Experiment #1 69

5.2 Experiment #2 86

5.3 Experiment #3 90

6 Conclusions and Discussion **93**

7 References **96**

List of Tables

3.1 Table with the fold changes of a subset of the Male/Female data that corresponds to the genes selected by LASSO. 20
3.2 Table with the coefficients of the regression model computed by LASSO using the Male/Female data. 21
3.3 Table with the fold changes of a subset of the Male/Female data that corresponds to the genes selected by SMLR. 24
3.4 Table with the coefficients of the regression model computed by SMLR using the Male/Female data. 25
3.5 Table with the number of all the genes selected by each algorithm using the Male/Female data. 26
3.6 Table with the fold changes of a subset of the Intersex/Female data that corresponds to the genes selected by LASSO. 30
3.7 Table with the coefficients of the regression model computed by LASSO using the Intersex/Female data. 31
3.8 Table with the fold changes of a subset of the Intersex/Female data that corresponds to the genes selected by SMLR. 34
3.9 Table with the coefficients of the regression model computed by SMLR using the Intersex/Female data. 35
3.10 Table with the number of all the genes selected by each algorithm using the Intersex/Female data. 36
3.11 Table with the fold changes of a subset of the Intersex/Male data that corresponds to the genes selected by LASSO. 40
3.12 Table with the coefficients of the regression model computed by LASSO using the Intersex/Male data. 41
3.13 Table with the fold changes of a subset of the Intersex/Male data that corresponds to the genes selected by SMLR. 44

3.14 Table with the coefficients of the regression model computed by SMLR using the Intersex/Male data. 45
3.15 Table with the number of all the genes selected by each algorithm using the Intersex/Male data. 46
3.16 Table with a summary of the number of all the genes selected by all algorithms in each experiment 46
4.1 Table with the fold changes of a subset of the ER+/ER- data that corresponds to the genes selected by LASSO. (ER+ only) 52
4.2 Table with the fold changes of a subset of the ER+/ER- data that corresponds to the genes selected by LASSO. (ER- only) 53
4.3 Table with the coefficients of the regression model computed by LASSO using the ER+/ER- data. 53
4.4 Table with the coefficients of the regression model computed by SLR using the ER+/ER- data. 55
4.5 Table with the number of all the genes selected by each algorithm using the ER+/ER- data. 56
4.6 Table with the fold changes of a subset of the LN-/LN+ data that corresponds to the genes selected by LASSO (LN+ only). 60
4.7 Table with the fold changes of a subset of the LN-/LN+ data that corresponds to the genes selected by LASSO (LN- only). 61
4.8 Table with the coefficients of the regression model computed by LASSO using the LN-/LN+ data. 62
4.9 Table with the number of all the genes selected by each algorithm using the LN-/LN+ data. 64
4.10 Table with the fold changes of a subset of the ER+LN+/ER+LN-/ER-LN+/ER-LN- data that corresponds to the genes selected by SMLR (ER-LN- and ER-/LN+ only). 66

4.11 Table with the fold changes of a subset of the ER+LN+/ER+LN-/ER-LN+/ER-LN- data that corresponds to the genes selected by SMLR (ER+/LN+ and ER+LN- only). 67
5.1 Table with the number of differentially expressed genes correctly and incorrectly chosen by LASSO using the simulated data with 5% regulated genes. 72
5.2 Table with the number of differentially expressed genes correctly and incorrectly chosen by SLR using the simulated data with 5% regulated genes. 74
5.3 Table with the number of differentially expressed genes correctly and incorrectly chosen by SMLR using the simulated data with 5% regulated genes. 75
5.4 Summary of the experiment that used the simulated data with 5% regulated genes. 77
5.5 Table with the number of differentially expressed genes correctly and incorrectly chosen by LASSO using the simulated data with 5% regulated genes. 80
5.6 Table with the number of differentially expressed genes correctly and incorrectly chosen by SLR using the simulated data with 5% regulated genes. 81
5.7 Table with the number of differentially expressed genes correctly and incorrectly chosen by SMLR using the simulated data with 5% regulated genes. 83
5.8 Summary of the experiment that used the simulated data with 15% regulated genes. 85
5.9 Summary of the experiment that used different # of training samples. 88
5.10 True/False positive rates of the experiments that used different # of training samples. 89
5.11 Percentage of correctly identifying to which conditions the 200 samples belong. 89
5.12 Summary of the experiment that used different # of training samples. 91
5.13 True/False positive rates of the experiment that used different # of training samples. 91
5.14 The percentage of correctly identifying to which condition a future sample belongs. 91

List of Figures

1.1 The general process of a single-colour microarray analysis. Source: http://www.nationalacademies.org/ 2
1.2 A digital image of the spots in the hybridized microarray. Source: http://www.nationalacademies.org/ 3
1.3 A microarray spot. Source: http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray 4
3.1 Boxplot of Male/Female log_2(Fold Change data) by replicate. 17
3.2 Scatter plot of Male/Female log_2(Fold Change data) by gene number. 18
3.3 Scatter plot of Male/Female log_2(Fold Change data) with genes selected by LASSO. 19
3.4 Scatter plot of Male/Female log_2(Fold Change data) with gene selected by SLR. 22
3.5 Scatter plot of Male/Female log_2(Fold Change data) with genes selected by SMLR. 23
3.6 Scatter plot of Male/Female log_2(Fold change data) of selected genes against their corresponding SMLR model coefficients. 26
3.7 Boxplot of Intersex/Female log_2(Fold Change data) by replicate. 27
3.8 Scatter plot of Intersex/Female log_2(Fold Change data) by gene number. 28
3.9 Scatter plot of Intersex/Female log_2(Fold Change data) with genes selected by LASSO. 29
3.10 Scatter plot of Intersex/Female log_2(Fold Change data) with genes selected by SLR. 32
3.11 Scatter plot of Intersex/Female log_2(Fold Change data) with genes selected by SMLR. 33
3.12 Scatter plot of Intersex/Female log_2(Fold change data) of selected genes against their corresponding SMLR model coefficients. 36
3.13 Boxplot of Intersex/Male log_2(Fold Change data) by replicate. 37
3.14 Scatter plot of Intersex/Male log_2(Fold Change data) by gene number. 38

3.15 Scatter plot of Intersex/Male log_2(Fold Change data) with genes selected by LASSO. 39
3.16 Scatter plot of Intersex/Male log_2(Fold Change data) with genes selected by SLR. 42
3.17 Scatter plot of Intersex/Male log_2(Fold Change data) with genes selected by SMLR. 43
3.18 Scatter plot of Intersex/Male log_2(Fold change data) of selected genes against their corresponding SMLR model coefficients. 45
4.1 Boxplot of ER+/ER- log_2(Fold Change data) by replicate. 49
4.2 Scatter plot of ER+/ER- log_2(Fold Change data) by gene number. 50
4.3 Scatter plot of ER+/ER- log_2(Fold Change data) with genes selected by LASSO. 51
4.4 Scatter plot of ER+/ER- log_2(Fold Change data) with genes selected by SLR. . 54
4.5 Scatter plot of ER+/ER- log_2(Fold Change data) with genes selected by SMLR. 56
4.6 Boxplot of LN-/LN+ log_2(Fold Change data) by replicate. 57
4.7 Scatter plot of LN-/LN+ log_2(Fold Change data) by gene number. 58
4.8 Scatter plot of LN-/LN+ log_2(Fold Change data) with genes selected by LASSO. 59
4.9 Scatter plot of LN-/LN+ log_2(Fold Change data) with genes selected by SLR. . 63
4.10 Scatter plot of LN-/LN+ log_2(Fold Change data) with genes selected by SMLR. 64
5.1 Boxplot of log_2(simulated Fold Change data) with 5% regulated genes by replicate. 70
5.2 Scatter plot of log_2(simulated Fold Change data) with 5% regulated genes by gene number. 71
5.3 Scatter plot of 5% regulated genes log_2(Fold Change data) with genes selected by LASSO. 72
5.4 Scatter plot of 5% regulated genes log_2(Fold Change data) with genes selected by SLR. 73
5.5 Scatter plot of 5% regulated genes log_2(Fold Change data) with genes selected by SMLR. 75
5.6 Scatter plot of 5% regulated genes log_2(Fold Change data) of selected genes against their corresponding SMLR model coefficients. 76

5.7 Boxplot of log_2(simulated Fold Change data) with 15% regulated genes by replicate. 78
5.8 Scatter plot of log_2(simulated Fold Change data) with 15% regulated genes by gene number. 79
5.9 Scatter plot of 15% regulated genes log_2(Fold Change data) with genes selected by LASSO. 80
5.10 Scatter plot of 15% regulated genes log_2(Fold Change data) with genes selected by SLR. 81
5.11 Scatter plot of 15% regulated genes log_2(Fold Change data) with genes selected by SMLR. 82
5.12 Scatter plot of 15% regulated genes log_2(Fold Change data) of selected genes against their corresponding SMLR model coefficients. 84

Acknowledgments

I would first like to express my deepest gratitude to my supervisor, Dr. Mary Lesperance. Her vast experience, enthusiasm towards research, and positive attitude have made it very inspiring to work under her guidance. I am deeply indebted to her for her encouragement and kind support, which have made this thesis possible.

Secondly, I am grateful to the members of the supervisory committee for their insightful comments and suggestions.

Lastly, and most importantly, I wish to thank my parents for always being there and for giving me all their support.

1 Introduction

1.1 Microarray

Functional genomics involves the analysis of huge datasets containing information derived from various biological experiments. Gene expression analysis is an example of a large-scale experiment, where one measures the transcription of the genetic information contained within the DNA into other products, for example, messenger RNA (mRNA). The mRNA is then translated into proteins, which in turn carry out most of the critical functions of cells. Gene expression is a sophisticated and closely regulated process. It differs in the kind and amount of mRNA production. By studying different levels of mRNA activities of a cell, scientists learn how the cell changes to respond both to environmental stimuli and its own needs. However, gene expression involves monitoring the expression levels of thousands of genes simultaneously under a particular condition. Microarray technology makes this possible. A microarray is a tool for analyzing gene expression. It consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. Microarray analysis allows scientists to detect thousands of genes in a small sample simultaneously and to analyze the expression of those genes. It may be used to examine gene expression within a single sample or to compare gene expression in two different conditions, such as in healthy and diseased tissue. For single-colour microarrays, each slide represents only one condition. To compare the gene expression of healthy and diseased tissue, two slides are needed. The general process of a single-colour microarray analysis is shown in Figure 1.1.

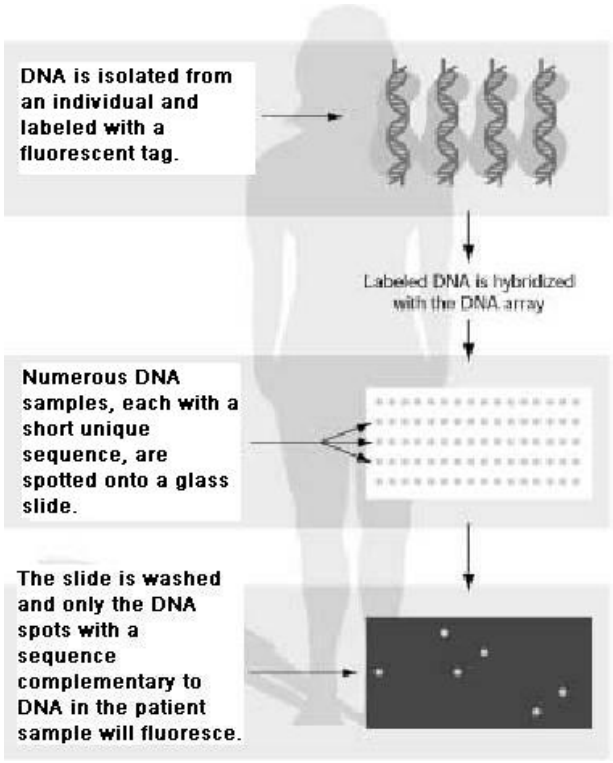

Figure 1.1: The general process of a single-colour microarray analysis. Source: http://www.nationalacademies.org/

The first step of the analysis is to spot numerous probe DNA samples onto a glass slide. Then isolate mRNA from a tissue sample, and use this mRNA as templates to generate cDNA with a fluorescent tag attached. A hybridization solution containing the fluorescently labeled cDNA is formed. The slide is washed with the hybridization solution and only the DNA spots with a sequence complementary to DNA in the sample will fluoresce. After this hybridization step, the spots in the hybridized microarray are excited by a laser and scanned. A digital image similar to Figure 1.2 is created. Each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene.

Figure 1.2: A digital image of the spots in the hybridized microarray. Source: http://www.nationalacademies.org/

The image is then analyzed by image processing software. There are many different image processing software packages on the market. Imagene is one of the popular ones. This software automatically identifies spots and distinguishes them from noise. For each spot, it also determines the spot area to be examined and determines the local region to estimate background hybridization. Once the spot and background areas have been defined, quantitative data is reported. The mean, median, and total values for the intensity considering all the pixels in the defined area, as in Figure 1.3, are reported for both the spot and background.

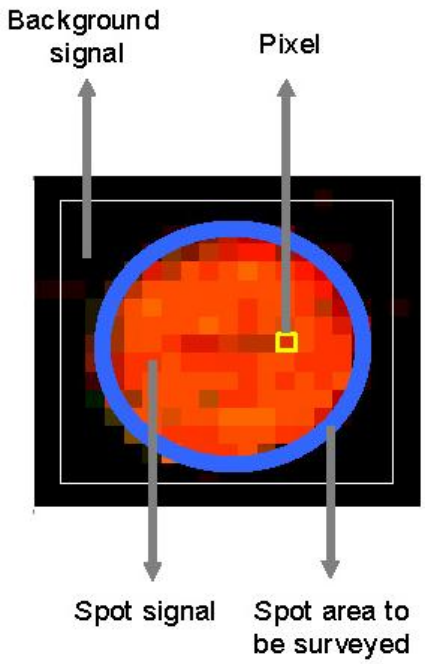

Figure 1.3: A microarray spot. Source: http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray

2 Literature Review

As mentioned in the previous chapter, a microarray is a tool for analyzing gene expression, which allows scientists to detect thousands of genes in a small sample simultaneously and to analyze the expression of those genes. This implies that a large dataset containing the measurements of the genes' expression levels will be produced. However, since a microarray is a costly experiment, few replications are usually done. Considering each gene as one variable, letting p represent the number of variables and n represent the number of replicates, a dataset with $p \gg n$ is generated. Datasets that have $p \gg n$ have always been a problem for statisticians, while methods such as dimension reduction and variable selection are often used to deal with such problems.

2.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Tibshirani (1995) proposed a technique called lasso, or 'least absolute shrinkage and selection operator'. This method fits a model to the data using a method similar to ordinary least squares (OLS), except that a constraint shrinks the linear model by setting some of the coefficients to zero.

Suppose that the data has a form (x_i, y_i), $i = 1, 2, ..., n$, where $x_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ are the predictor variables and y_i are the responses. Observations are assumed to be independent as in a usual regression. It is also assumed that x_{ij}'s are standardized with mean zero and variance one. Letting $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$, the lasso estimate $\hat{\beta}$ is defined by

$$\hat{\beta} = \arg\min \left[\sum_{i=1}^{n} (y_i - \sum_j \beta_j x_{ij})^2 \right] \quad \text{such that} \sum_{j=1}^{p} |\beta_j| \leq t. \tag{2.1.1}$$

where $t \geq 0$ is a tuning parameter that controls the amount of linear model shrinkage, in other words, the number of estimates that are set to zero. It is suggested in the article that if the full least squares estimates is $\hat{\beta}_j^o$, then one should set $t_0 = \sum |\hat{\beta}_j^o|$. When $t < t_0$, the method will shrink the model, forcing some of the estimates to approach zero. For example, the effect

of letting $t = t_0/2$ will have similar effect as finding the best subset of estimates with size $p/2$.

Tibshirani (1995) claimed in his study that lasso works better in improving OLS estimates compared to both subset selection and ridge regression. The algorithm not only can increase the prediction accuracy of OLS estimates, but also offers an interpretable model at the same time. In addition, Tibshirani (1995) suggested that further extensions can be made from lasso to fit different types of data. The Lasso routine is in the R package (R Development Core Team, 2008), lasso2, and is called l1ce. This function is used in this study.

2.2 Sparse Logistic Regression (SLR)

Shevade and Keerthi (2003) later proposed another technique following Tibshirani's suggestion. Lasso is extended to a generalized regression model, which is named Sparse Logistic Regression. Instead of fitting a linear model, logistic regression is used. Estimates are obtained by making use of Maximum Likelihood Estimation. The negative log-likelihood function of the binomial distribution is minimized to obtain the estimates. A constraint that will shrink the logistic regression model, which is similar to the one in Tibshirani's model, is applied. Only response variables with two categories can be analyzed using this technique.

As for lasso, suppose that the data are (x_i, y_i), $i = 1, 2, ..., n$, where $x_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ are the predictor variables. Now, y_i is the response variable which takes on values 1 or -1; $y_i = 1$ means x_i is in class 1 and $y_i = -1$ means x_i is in class 2. Letting $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$, the Sparse Logistic Regression problem is formulated as follows:

$$\begin{aligned} \hat{\beta} &= \min_{\beta} \sum_{i} g\left[-y_i f(x_i)\right] \\ &\text{such that} \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{2.2.1}$$

where $t \geq 0$ is again a tuning parameter, which has a similar function as the one in Tibshirani's (1995) model. The regression function $f(x)$ has the form of a linear model,

$$f(x_i) = \sum_{j=0}^{p} \beta_j x_{ij}. \tag{2.2.2}$$

The function g is given by:

$$g(\xi) = \log(1 + e^{\xi}) \tag{2.2.3}$$

This is the negative log-likelihood function associated with the probabilistic model

$$\text{Prob}(y|x) = \frac{1}{1 + e^{-yf(x)}} \tag{2.2.4}$$

Shevade and Keerthi (2003) suggested that this algorithm not only works as a variable selection tool, it also functions as a classifier which separates tissue samples into the two predefined classes. Once the β's are determined by solving (2.2.1), the class of the test sample, $\tilde{x}$, belongs to class 1 if $f(\tilde{x}) > 0$ and belongs to class 2 otherwise. Shevade and Keerthi (2003) proposed an algorithm for Sparse Logistic Regression in their article that neither uses any mathematical programming package nor needs any matrix operations. This algorithm, with minor changes, is programmed in this study using R.

In (2.2.1), the negative log-likelihood function of a logistic regression model is minimized to compute the Maximum Likelihood Estimates (MLE) of the regression coefficients. Shevade and Keerthi (2003) suggested the existence of a $\gamma > 0$, for which (2.2.1) is equivalent to the following unconstrained optimization problem:

$$\min_{\beta} W = \gamma \sum_{j=1}^{p} |\beta_j| + \sum_{i} g\left[-y_i f(x_i)\right] \tag{2.2.5}$$

The article indicates that the family of classifiers obtained by varying t in (2.2.1) and the family obtained by varying γ in (2.2.5) are the same. The method proposed by Shevade and Keerthi (2003) optimizes one variable at a time while keeping the other variables fixed. This process is repeated until the optimality conditions are satisfied. Shevade and Keerthi (2003) also state that the strict convexity characteristic of the g function in the negative log-likelihood function of (2.2.1) plays an important role in this algorithm. Before introducing the first order

optimality conditions for (2.2.5), let us define

$$\begin{aligned} \xi_i &= -y_i f(x_i) \\ F_j &= \sum_i \frac{e^{\xi_i}}{1+e^{\xi_i}} y_i x_{ij} \end{aligned} \tag{2.2.6}$$

Because of the absolute value in the first part of (2.2.5), there are some points where W is not differentiable. The first order optimality conditions are given in Shevade and Keerthi (2003) as follows:

1. Since W is differentiable with respect to β_0, $\frac{\partial W}{\partial \beta_0} = 0$;
2. For $j > 0$ and $\beta_j \neq 0$, since W is differentiable with respect to β_j, for $\beta_j \neq 0$, $\frac{\partial W}{\partial \beta_j} = 0$
3. For $j > 0$ and $\beta_j = 0$, W is only directionally differentiable with respect to $\beta_j = 0$. It is required that the right side derivative of W with respect to β_j to be non-negative and the left side derivative to be non-positive.

The three conditions are defined algebraically as follows:

$$\begin{aligned} F_j &= 0 && \text{if } j = 0 \\ F_j &= \gamma && \text{if } \beta_j > 0, j > 0 \\ F_j &= -\gamma && \text{if } \beta_j < 0, j > 0 \\ -\gamma \leq F_j &\leq \gamma && \text{if } \beta_j = 0, j > 0 \end{aligned}$$

This can also be written as

$$viol_j = \begin{cases} |F_j|, & \text{if } j = 0 \\ |\gamma - F_j|, & \text{if } \beta_j > 0, j > 0 \\ |\gamma + F_j|, & \text{if } \beta_j < 0, j > 0 \\ \psi_j, & \text{if } \beta_j = 0, j > 0 \end{cases} \tag{2.2.7}$$

where $\psi_j = \max(F_j - \gamma, -\gamma - F_j, 0)$. Then the first order conditions can be simply written as:

$$viol_j = 0 \quad \text{for all } j, \tag{2.2.8}$$

Because in finite time, it is difficult to achieve exact optimality in an asymptotically convergent procedures, they assume convergence when a tolerance, τ, is achieved. In the SLR case, the algorithm is terminated when

$$viol_j \leq \tau \text{ for all } j, \text{ and a small } \tau > 0 \tag{2.2.9}$$

This SLR algorithm developed by Shevade and Keerthi (2003) uses the Gauss-Seidel method (Hageman and Young, 1981). One variable β_j that violates the optimality condition is first chosen and the objective function (2.2.5) is optimized with respect to this variable β_j, while keeping the other β_j's fixed. This procedure is repeated until there are no variables violating the optimality conditions. Shevade and Keerthi (2003) state that the objective function (2.2.5) is strictly convex, which implies that it will strictly decrease at every step. To explain the details of this algorithm, the following sets are defined: $I_z = \{j : \beta_j = 0, j > 0\}$; and $I_{nz} = \{0\} \cup \{j : \beta_j \neq 0, j > 0\}$. Also, let $I = I_z \cup I_{nz}$. The procedure is explained as follows:

Input Training Examples

Initialize β's to 0

While an Optimality Violator exists in I_z

Find the maximum violator, ν, in I_z,

(i.e. the β_ν such that $viol_\nu$ is a maximum)

Repeat

Optimize W with respect to β_ν

Find the maximum violator, ν , in I_{nz}

Until No violator exists in I_{nz}

End while

Output A set of β's for the function in (2.2.1) .

The algorithm continues until no violators exist in either of the sets, I_z and I_{nz}. In the algorithm, once the maximum violator is chosen, some non-linear optimization is required to solve the problem (2.2.5) with respect to one variable. Note that the objective function (2.2.5) is convex. In the article, Shevade and Keerthi (2003) used a combination of the bisection method and Newton's method to solve the optimization problem. The Bisection method is used to locate the interval depending on the signs the derivative of the objective function, while the Newton's method is restricted to locate the optimum within this interval. In this study, a similar approach is employed. Instead of using the Newton's method, a built-in function in R, **Optimize**, is used to locate the optimum.

2.3 Sparse Multinomial Logistic Regression (SMLR)

While Sparse Logistic Regression can only work with data in two categories, a further extension that is based on the multinomial logistic regression is developed by Krishnapuram et al. (2005), and is named Sparse Multinomial Logistic Regression. This method deals with data having more than two categories.

Suppose that there is data (x_k, y_k), $k = 1, 2, ...n$, where $x_k = (x_{k1}, x_{k2}..., x_{kp})$ are the predictor variables. Krishnapuram et al. (2005) adopt the technique of representing the class labels using indicators of category membership, $y_k = [y_k^{(1)}, y_k^{(2)}, ..., y_k^{(m)}]$, where m is the number of categories. In this case, $y_k^{(i)} = 1$ if the k'th observation belongs to class i and $y_k^{(i)} = 0$ otherwise. Under the multinomial logistic regression model, the probability that the k'th observation belongs to class i can be formulated as

$$P(y_k^{(i)} = 1|x_k, \omega) = \frac{e^{\omega^{(i)\prime} x_k}}{\sum_{j=1}^{m} e^{\omega^{(j)\prime} x_k}}, \tag{2.3.1}$$

where $\omega^{(i)}$ is the vector parameter corresponding to class i. For the problem with only two categories (m=2), this is same as a logistic regression model. Given the constraint that

$$\sum_{i=1}^{m} P(y_k^{(i)} = 1|x_k, \omega) = 1,$$

one of the $\omega^{(i)}$ can be phrased in terms of the others. Krishnapuram et al. (2005) suggested that, without the loss of generality, the weight of the last category, $\omega^{(m)}$, is set to zero; therefore, only $m-1$ vectors will be estimated. In their article, ω was used to denote the $p(m-1)$-dimensional vector of parameters to be estimated, and the same notation is used in this study. Similar to SLR, the maximum likelihood estimation method is used to estimate the parameters. The log-likelihood function associated with probabilistic model is formulated as

$$\begin{aligned} l(\omega) &= \sum_{k=1}^{n} \log P(y_k|x_k,\omega) \\ &= \sum_{k=1}^{n} \left[\sum_{i=1}^{m} y_k^{(i)} \omega^{(i)'} x_k - \log \sum_{i=1}^{m} e^{\omega^{(i)'} x_k} \right]. \end{aligned} \tag{2.3.2}$$

Krishnapuram et al. (2005) suggested that $l(\omega)$ can be large depending on the data; therefore, a prior on ω was introduced to restrict the log-likelihood function. This motivated their employment of the penalized maximum likelihood estimate

$$\hat{\omega} = \arg\max_{\omega} L(\omega) = \arg\max_{\omega} \left[l(\omega) + \log p(\omega) \right]$$

with $p(\omega)$ being some prior on ω . One of the priors introduced in the paper, which is called the Laplacian prior, contributed to formulate a method that is very similar to the one established in SLR:

$$p(\omega) \propto e^{-\lambda \|\omega\|_1} \tag{2.3.3}$$

where $\| \omega \|_1 = \sum_l |\omega_l|$. After incorporating the Laplacian Prior, the objective function $L(\omega)$ becomes non-differentiable at its origin. Krishnapuram et al. (2005) described in their article a bound optimization algorithm to estimate ω that can be used with a Laplacian prior.

Krishnapuram et al. (2005) use the bound optimization approach, that is, the objective function $L(\omega)$ is optimized by iteratively maximizing a surrogate function Q,

$$\hat{\omega}^{(t+1)} = \arg\max_{\omega} Q(\omega|\hat{\omega}^{(t)}). \tag{2.3.4}$$

With this procedure, the objective function will increase at each iteration if the surrogate

function satisfies the condition that $L(\omega) - Q(\omega|\hat{\omega}^{(t)})$ attains its minimum when $\omega = \hat{\omega}^{(t)}$. This is shown by following:

$$\begin{aligned} L(\hat{\omega}^{(t+1)}) &= L(\hat{\omega}^{(t+1)}) - Q(\hat{\omega}^{(t+1)}|\hat{\omega}^{(t)}) + Q(\hat{\omega}^{(t+1)}|\hat{\omega}^{(t)}) \\ &\geq L(\hat{\omega}^{(t)}) - Q(\hat{\omega}^{(t)}|\hat{\omega}^{(t)}) + Q(\hat{\omega}^{(t+1)}|\hat{\omega}^{(t)}) \\ &\geq L(\hat{\omega}^{(t)}) - Q(\hat{\omega}^{(t)}|\hat{\omega}^{(t)}) + Q(\hat{\omega}^{(t)}|\hat{\omega}^{(t)}) \\ &= L(\hat{\omega}^{(t)}) \end{aligned} \tag{2.3.5}$$

The condition that $L(\omega) - Q(\omega|\hat{\omega}^{(t)})$ attains its minimum when $\omega = \hat{\omega}^{(t)}$ leads to the first inequality of (2.3.5) , while the fact that $Q(\omega|\hat{\omega}^{(t)})$ is maximized when $\omega = \hat{\omega}^{(t+1)}$ contributes to the second inequality. Krishnapuram et al. (2005) suggest in their article a way to obtain a surrogate function $Q(\omega|\omega')$ when the objective function $L(\omega)$ is concave. They use a bound on the Hessian $L(\omega)$. If there exists a negative definite matrix B such that $H(\omega) \geq B$, with the Taylor series expansion, it is simple to show that, for any ω^*,

$$L(\omega) \geq L(\omega^*) + (\omega - \omega^*)'g(\omega^*) + \frac{1}{2}(\omega - \omega^*)'B(\omega - \omega^*) \tag{2.3.6}$$

where $g(\omega^*)$ denotes the gradient function of $L(\omega)$ at ω^*. If we let the right side of the above inequality be the surrogate function $Q(\omega|\omega^*)$, it can be rewritten to $L(\omega) - Q(\omega|\omega^*) \geq 0$, with equality if and only if $\omega = \omega^*$. This implies that $Q(\omega|\omega^*)$ here is a valid surrogate function. With the terms that are irrelevant for the optimization being removed, a monotonic function is obtained.

$$Q(\omega|\hat{\omega}^{(t)}) = \omega' \left(g(\hat{\omega}^{(t)}) - B\hat{\omega}^{(t)} \right) + \frac{1}{2}\omega' B\omega \tag{2.3.7}$$

The maximization of the surrogate function leads to the following update equation.

$$\hat{\omega}^{(t+1)} = \hat{\omega}^{(t)} - B^{-1}g(\hat{\omega}^{(t)}). \tag{2.3.8}$$

This bound optimization algorithm was then applied to maximum likelihood multinomial logistic regression. Let $p_j^{(i)}(\omega) = P(y_j^{(i)} = 1|x_j, \omega)$, the vector $p_j(\omega) = [p_j^{(1)}(\omega), ..., p_j^{(m-1)}(\omega)]'$,

and the diagonal matrix $P_j(\omega) = diag\{p_j^{(1)}(\omega), ..., p_j^{(m-1)}(\omega)\}$. For multinomial logistic regression, $L(\omega)$ is just the log-likelihood $l(\omega)$ given in (2.3.2). This is a concave function with Hessian

$$H(\omega) = -\sum_{j=1}^{n} \left[P_j(\omega) - p_j(\omega)p_j'(\omega)\right] \otimes (x_j x_j'), \tag{2.3.9}$$

where $\otimes$ is the Kronecker matrix product. Böhning (1992) shows that the Hessian is bounded by a negative definite matrix that does not depend on ω,

$$H(\omega) \geq -\frac{1}{2}[\mathbf{I} - \mathbf{1}\mathbf{1}'/m] \otimes \sum_{j=1}^{n} x_j x_j' \equiv B \tag{2.3.10}$$

Where $\mathbf{I}$ is an identity matrix and $\mathbf{1} = [1, 1..., 1]'$. The gradient of $l(\omega)$ has a closed form:

$$g(\omega) = \sum_{j=1}^{n} \left(y_j' - p_j(\omega)\right) \otimes x_j \tag{2.3.11}$$

where $y_j' = [y_j^{(1)}, y_j^{(2)}, ..., y_j^{(m-1)}]'$.

The bound optimization algorithm was further extended to apply not only to the multinomial logistic regression, but also to the problem including the Laplacian prior. The only modification of the algorithm is that, in each iteration, we now have to maximize

$$Q(\omega|\hat{\omega}^{(t)}) - \lambda \parallel \omega \parallel_1 \tag{2.3.12}$$

which is equivalent to maximizing

$$\omega' \left(g(\hat{\omega}^{(t)}) - B\hat{\omega}^{(t)}\right) + \frac{1}{2}\omega' B\omega - \lambda \parallel \omega \parallel_1 \tag{2.3.13}$$

Krishnapuram et al. (2005) suggest that it may be problematic when p is very large; therefore, they introduced a component-wise update procedure to address the problem. The main idea is to maximize the surrogate function with respect to one of the components of ω, while keeping

the others constant. The component-wise update is formulated as

$$\hat{\omega}_k^{(t+1)} = soft\left(\hat{\omega}^{(t)} - \frac{g_k(\hat{\omega}^{(t)})}{B_{kk}}; \frac{-\lambda}{B_{kk}}\right)$$

where B_{ij} is the (i, j) element of matrix B, $g_k(\hat{\omega}^{(t)})$ is the kth element of the gradient vector, and

$$soft(a; \delta) = sign(a)max\{0, |a| - \delta\}$$

is the soft threshold function.

2.4 Model Setup (LASSO, SLR, and SMLR)

In the following chapters, LASSO, SLR, and SMLR are tested using both real and simulated gene data. A brief description on how the gene data was set up to test these alogorithms is provided in this section.

Beginning with LASSO, gene fold change data are fitted using Equation 2.1.1. It is suggested that the data processed by LASSO should have a form (x_i, y_i), $i = 1, 2, ..., n$, where $x_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ are the predictor variables and y_i are the responses. In the gene data context, x_i refers to a vector of fold changes of all genes in sample i, where p refers to the total number of genes in a sample, and n refers to the total number of samples. The response variable y_i refers to an indicator identifying to which condition sample i belongs. $y_i = 0$ when sample i belongs to the control condition, and $y_i = 1$ when sample i belongs to the treatment condition.

SLR has a setup similar to LASSO. When gene fold change data is fitted using Equation 2.2.1, the data should have a form (x_i, y_i), $i = 1, 2, ..., n$, where $x_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ are the predictor variables, and y_i are the responses. The same notation is used in SLR, that x_i refers to a vector of fold changes of all genes in sample i, where p refers to the total number of genes in a sample, and n refers to the total number of samples. In this algorithm, the response variable y_i still refers to an indicator identifying to which condition sample i belongs, but the values are different. In this case $y_i = -1$ when sample i belongs to the control condition, and $y_i = 1$ when sample i belongs to the treatment condition.

When fitting the data using Equation 2.3.2 with SMLR, the gene data should have a form (x_k, y_k), $k = 1, 2, ...n$, where $x_k = (x_{k1}, x_{k2}..., x_{kp})$ are the predictor variables, and y_k are the responses. x_k refers to a vector of fold changes of all genes in sample i, where p refers to the total number of genes in a sample, and n refers to the total number of samples. y_k here is a $m \times 1$ vector that indicates to which condition sample k belongs. Since SMLR is capable of handling more than two conditions, y_k is defined as $y_k = [y_k^{(1)}, y_k^{(2)}, ..., y_k^{(m)}]$, where m is the total number of conditions. In this case, $y_k^{(i)} = 1$ if the k'th observation belongs to class i and $y_k^{(i)} = 0$ otherwise.

Attention was paid to the fact that many researchers suggested transforming genetic data using a base-2 logarithm transformation before performing analyses, however, this approach did not work well when testing LASSO, SLR, and SMLR using our datasets. In many cases, the algorithms, especially SLR, did not not converge no matter what tuning parameter used. Therefore, fold change data was not transformed in our experiments, however, to allow a better view of the gene data, all figures were plotted with base-2 logarithm transformed fold changes.

3 Intersex Data Experiment

Two real datasets were used to compare the results from the three algorithms. The first dataset came from a collaboration between Dr. Caren Helbing from the University of Victoria, and Krista McCoy and Dr. Lou Guillette from the University of Florida, Gainesville. This experiment was initiated with a background of toxicology prediction. Krista McCoy collected Bufo marinus frogs from sites with intensive agriculture, which is more pesticide contaminated, and also from sites with no agriculture. She knew that there is an increased incidence of intersex, that is, individuals with testes and ovaries in the same gonad, in the more contaminated sites. Dr. Caren Helbing compared histologically normal male gonads to histologically normal females to histologically severe intersex animals, each with five replicates.

Since LASSO, and SLR can only accommodate data with two categories, the three conditions of the Intersex data were separated into three pairs (male/female, intersex/female, intersex/male), and were transformed from intensities into fold changes. The fold changes were computed by dividing the intensities of each replicate of both sex groups by the average of intensities of all replicates of one of the two sex groups. Each pair of data was then used to test and compare the ability of the three different algorithms in selecting feature genes. The complete Intersex data was also employed later to examine SMLR's ability in working with data that has more than two categories.

3.1 Male/Female Experiment

The first pair of the Intersex data examined using the three algorithms was the male and female pair. The fold changes were computed by dividing the replicates from both sex groups by the average of the replicates from the female group. Figure 3.1 below provides a description of the data. The base-2 logarithm transformed fold changes of all genes of all replicates in both sex groups were included in this boxplot.

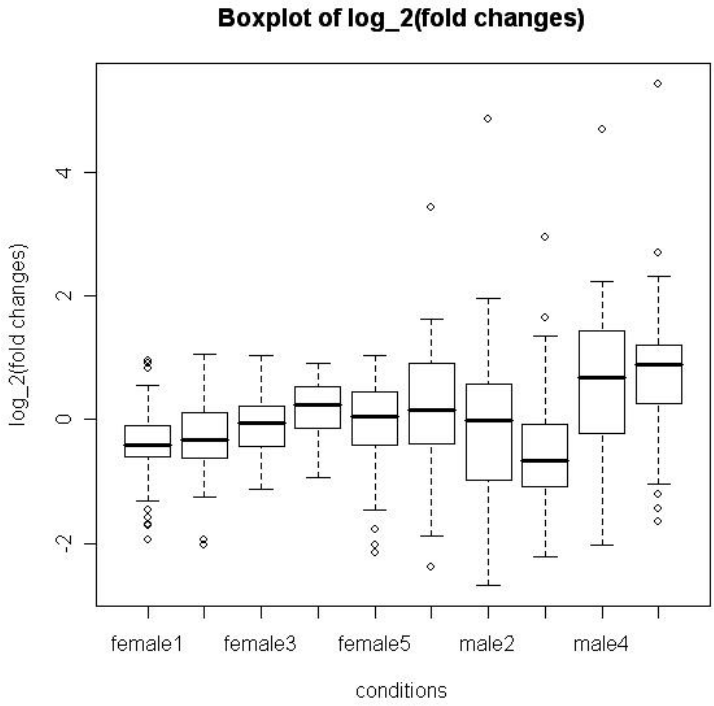

Figure 3.1: Boxplot of Male/Female log_2(Fold Change data) by replicate.

From the boxplot, we see that there is large variation within each of the male and the female samples. Also, it is clear that there are a few genes that contribute to great variability when comparing the genes from the two sex groups. This is evidenced by the outlying dots for the male boxplots which indicate large male/female ratios. Figure 3.2 provides a better look at this gene data.

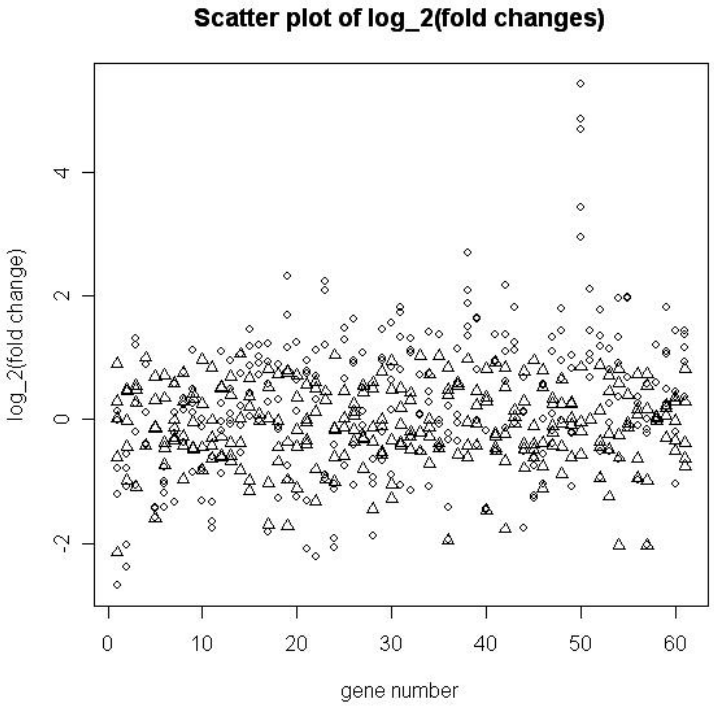

Figure 3.2: Scatter plot of Male/Female log_2(Fold Change data) by gene number.

Figure 3.2 shows the base-2 logarithm transformed fold changes of each gene from the ten replicates; five replicates from each sex group. The triangles represent the fold changes from the female group, while the circles represent those from the male group. From this graph, we see that the gene numbered 50 was responsible for the most variability, while there were other genes that act quite differently when compared between the individuals from the two sex groups, for example, gene numbers 19, 23, and 38. The data was processed by each of the three algorithms to determine whether they select the genes that are responsible for the most variation.

The first algorithm tested was LASSO. There is only a range of tuning parameter values that allows the algorithm to stay converged. Many tuning parameter values were tried, and the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 0.8 was used in this experiment. With this tuning parameter, nine genes were chosen. Figure 3.3 is a scatter plot similar to Figure 3.2 with the genes selected plotted in red.

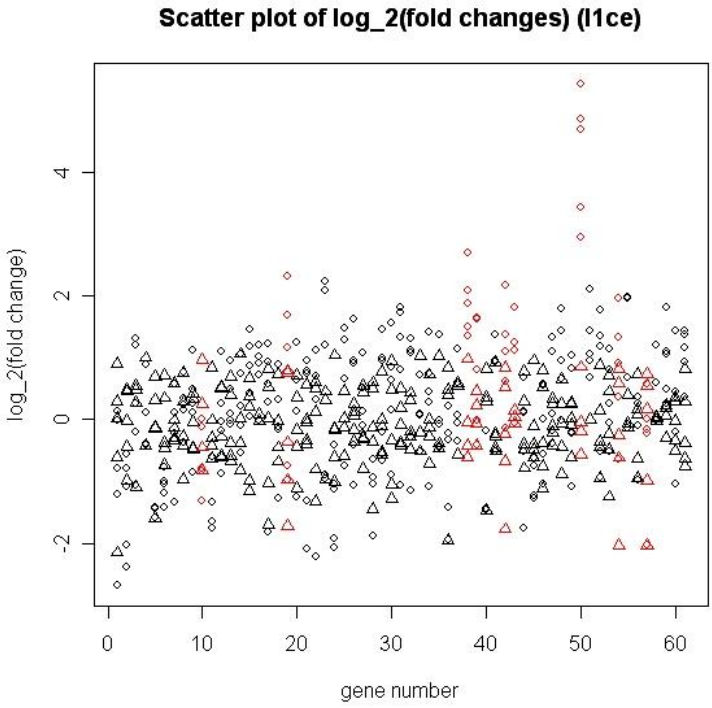

Figure 3.3: Scatter plot of Male/Female log_2(Fold Change data) with genes selected by LASSO.

From this plot, we see that gene numbered 50 was selected. This is the gene that displayed large variability in the boxplot (Figure 3.1). In addition, some other genes were selected by LASSO. These are the ones with larger fold changes that were obviously shown in Figure 3.2. Since the intersex data is not a very big dataset, the fold changes of a subset of the fold change data that corresponds to the selected genes is provided in Table 3.1.

Gene #	Gene name	Female1	Female2	Female3	Female4	Female5	Median
10	FJ2	1.95	0.57	0.57	0.73	1.19	0.73
19	FM24	0.30	1.69	0.78	1.73	0.51	0.78
38	FM68	0.66	0.66	1.97	0.75	0.97	0.75
39	FN32	1.37	0.96	0.75	1.18	0.75	0.96
42	FI11	0.63	1.44	0.85	1.79	0.29	0.85
43	FI28Asm	1.10	0.96	0.96	0.96	1.03	0.96
50	FM40	0.67	1.80	0.67	0.88	0.97	0.88
54	FM3	0.84	1.76	0.65	1.51	0.24	0.84
57	FG7	1.47	0.24	1.12	1.66	0.50	1.12

Gene #	Gene name	Male1	Male2	Male3	Male4	Male5	Median
10	FJ2	0.40	0.56	0.92	0.58	1.00	0.58
19	FM24	2.23	0.50	0.60	3.20	4.95	2.23
38	FM68	2.82	3.68	2.55	4.27	6.47	3.68
39	FN32	3.07	0.98	3.14	0.75	1.75	1.75
42	FI11	2.60	2.14	0.90	4.48	1.54	2.14
43	FI28Asm	0.96	0.96	2.17	3.53	2.35	2.17
50	FM40	10.79	28.89	7.68	25.97	43.07	25.97
54	FM3	1.88	1.29	0.64	2.53	3.90	1.88
57	FG7	1.48	0.24	0.89	1.00	0.86	0.89

Table 3.1: Table with the fold changes of a subset of the Male/Female data that corresponds to the genes selected by LASSO.

While looking at the median of all the replicates for the nine selected genes shown in Table 3.1, there is no male down-regulated (fold change<0.5) gene. However, there are five male genes that are up-regulated (fold change>2). We may suspect that LASSO is more sensitive to up-regulated genes.

Table 3.2 below provides the coefficients of the regression model computed by LASSO. After comparing these coefficients to the fold changes of the genes chosen by LASSO given in Table 3.1, there seems to be no obvious relationship between the magnitude of the fold changes and their corresponding coefficients.

Gene #	Gene name	Coef
	Intercept	-0.255
10	FJ2	-0.013
19	FM24	-0.018
38	FM68	0.028
39	FN32	0.299
42	FI11	0.158
43	FI28Asm	0.023
50	FM40	0.026
54	FM3	-0.174
57	FG7	-0.062

Table 3.2: Table with the coefficients of the regression model computed by LASSO using the Male/Female data.

The next algorithm we tested using this data was SLR. Again, there is only a range of tuning parameter values that allows the algorithm to stay converged. A wide range of tuning parameter values were tried, and the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 0.5 was used in this experiment. It is interesting to see that only one gene was chosen by SLR. Figure 3.4 below shows the gene selected in red.

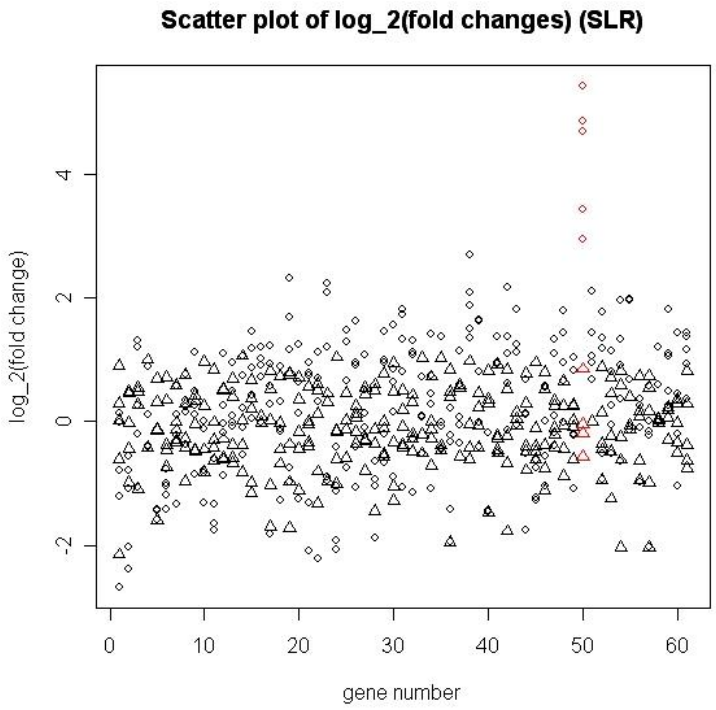

Figure 3.4: Scatter plot of Male/Female log_2(Fold Change data) with gene selected by SLR.

The only gene selected by SLR is a member of the list of genes selected by LASSO. It is the gene numbered 50, which is the gene that varied the most when compared between the genes obtained from the two sex groups. Since only one gene was chosen by the algorithm, it is impossible to comment on its sensitivity to up-regulated or down-regulated genes. However, we may suspect that the algorithm is not very sensitive to its tuning parameter when there is a gene in the data that is obviously different from the others. When this data was processed with SLR, various tuning parameter values were tried, but the results remained identical with only one gene selected. It seems that this algorithm only picked up the gene that contributed to great variation, while neglecting the others that might be meaningful but with comparatively less variation. A further investigation of the algorithm will be provided in the following sections.

Finally, the data was processed with the algorithm SMLR. As for LASSO and SLR, there is only a range of tuning parameters that allows the algorithm to stay converged. Different

tuning parameter values were tried and the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 0.32 was used in this experiment, and 10 genes were chosen by the algorithm. Figure 3.5 below shows the genes chosen in red.

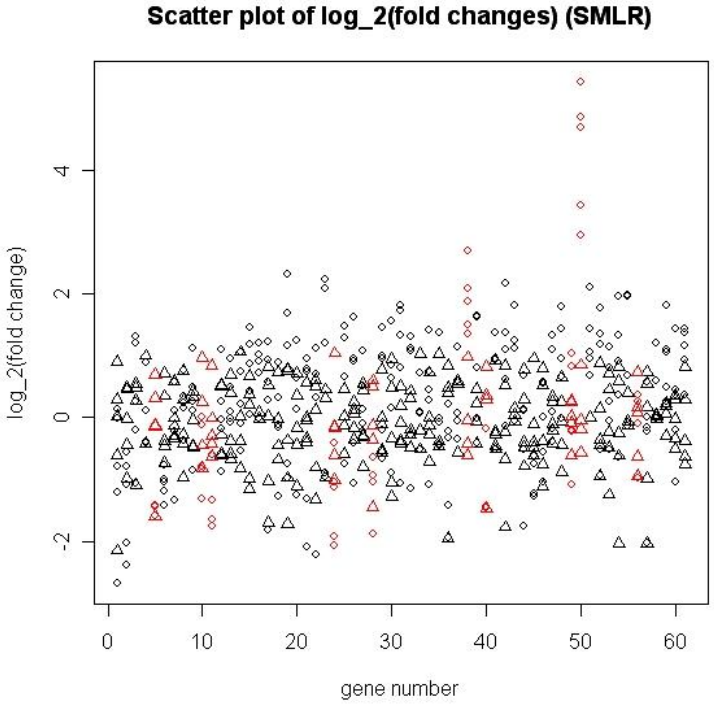

Figure 3.5: Scatter plot of Male/Female log_2(Fold Change data) with genes selected by SMLR.

It is not a surprise to see that the gene numbered 50 was again selected by the algorithm. The set of genes chosen by this algorithm is different from the ones chosen by LASSO. Similar to what was found with LASSO, SMLR selected the genes that are obviously different when comparing the samples from the two sex groups.

Gene #	Gene name	Female1	Female2	Female3	Female4	Female5	Median
5	FJ36	0.33	0.92	1.24	1.61	0.90	0.92
10	FJ2	1.95	0.57	0.57	0.73	1.19	0.73
11	FM62	0.82	0.75	0.65	0.99	1.79	0.82
24	FC6	0.89	0.50	0.65	0.90	2.05	0.89
28	FD3	0.37	1.42	1.51	0.92	0.78	0.92
38	FM68	0.66	0.66	1.97	0.75	0.97	0.75
40	FN44	0.36	1.76	1.29	1.23	0.36	1.23
49	FM69	1.01	0.95	1.20	1.18	0.65	1.01
50	FM40	0.67	1.80	0.67	0.88	0.97	0.88
56	FN19	1.07	1.65	1.12	0.52	0.64	1.07

Gene #	Gene name	Male1	Male2	Male3	Male4	Male5	Median
5	FJ36	0.33	0.37	0.37	0.33	0.37	0.37
10	FJ2	0.40	0.56	0.92	0.58	1.00	0.58
11	FM62	0.32	0.40	0.30	0.66	0.32	0.32
24	FC6	0.48	0.26	0.24	0.46	0.75	0.46
28	FD3	0.27	0.27	0.52	0.49	0.64	0.49
38	FM68	2.82	3.68	2.55	4.27	6.47	3.68
40	FN44	0.36	0.37	0.36	0.36	0.89	0.36
49	FM69	0.87	0.85	0.47	2.05	1.78	0.87
50	FM40	10.79	28.89	7.68	25.97	43.07	25.97
56	FN19	1.28	0.51	0.51	0.93	1.19	0.93

Table 3.3: Table with the fold changes of a subset of the Male/Female data that corresponds to the genes selected by SMLR.

From the the median column of Table 3.3, it is shown that among the ten genes selected by SMLR, only two of them are up-regulated genes. These are the obvious ones shown in Figure 3.2. Five of the genes selected are down-regulated. It is reasonable to suspect that this algorithm is more sensitive to down-regulated genes. Table 3.4 provides the coefficients of the multinomial logistic regression model computed using SMLR.

Gene #	Gene name	Coef
5	FJ36	0.384
10	FJ2	0.316
11	FM62	0.750
24	FC6	0.342
28	FD3	0.399
38	FM68	-0.210
40	FN44	0.585
49	FM69	0.066
50	FM40	-0.330
56	FN19	0.142

Table 3.4: Table with the coefficients of the regression model computed by SMLR using the Male/Female data.

When comparing the coefficients with the fold changes given in Table 3.3, a pattern on how SMLR determined coefficients is found. SMLR defined coefficients that correspond to up-regulated genes as negative numbers, and the opposite for the down-regulated genes. The magnitude of the coefficient for each gene was also decided based on the fold change of the gene. The magnitude of the coefficient was larger when the gene varied more between samples from the two sex groups. This applied to both up and down-regulated genes.

It is more easily visualized in Figure 3.6 below. In this figure, fold changes of the genes selected by SMLR were plotted against their corresponding coefficients. Genes with log_2(fold change) greater than 1 (up-regulated) had negative corresponding coefficients, while the genes with log_2(fold change) less than -1 (down-regulated) had positive coefficients. Also, the magnitude of the corresponding coefficients are closer to 0 with non-regulated genes (-1<log_2(fold change)<1).

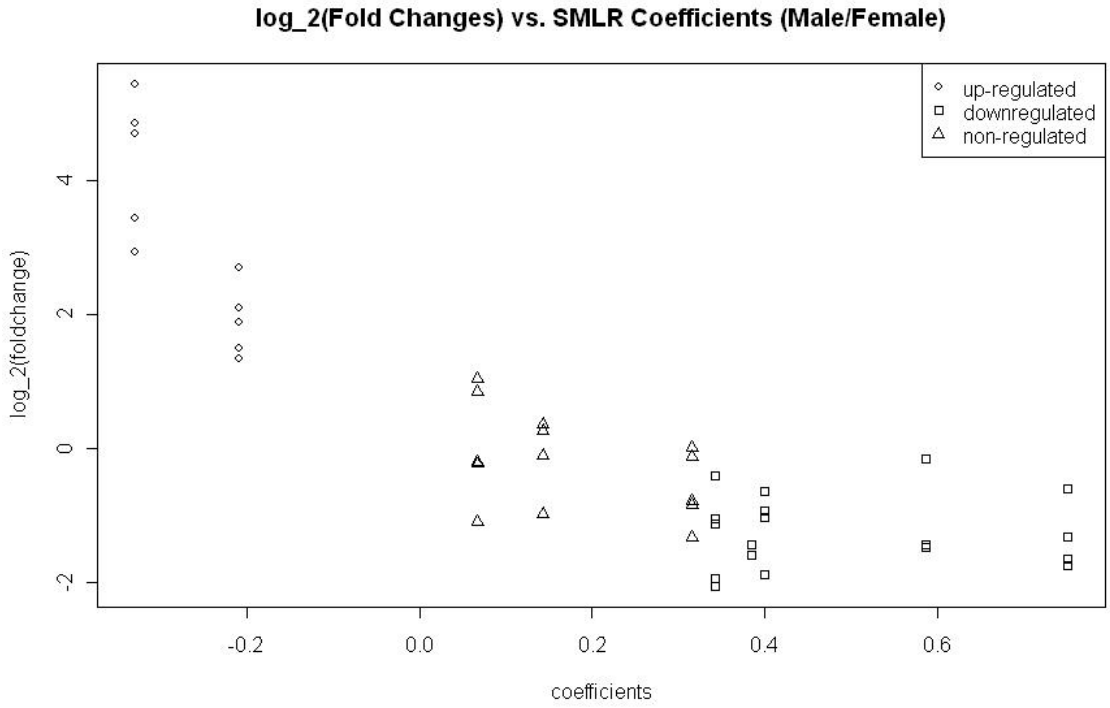

Figure 3.6: Scatter plot of Male/Female log_2(Fold change data) of selected genes against their corresponding SMLR model coefficients.

Algorithm	gene #
LASSO	10 19 38 39 42 43 50 54 57
SLR	50
SMLR	5 10 11 24 28 38 40 49 50 56

Table 3.5: Table with the number of all the genes selected by each algorithm using the Male/Female data.

Table 3.5 shows the number of all the genes selected by each algorithm. There is only one gene that was chosen by all of the three algorithms, while there are three of them that were selected by both LASSO and SMLR.

3.2 Intersex/Female Experiment

Another pair of sex groups from the Intersex data that was examined using the three algorithms was the female and intersex pair. The fold changes were computed by dividing the replicates

from both sex groups by the average of the replicates from the female group. Figure 3.7 provides a description of the data. The log_2(fold changes) of all genes from each replicate were included in this boxplot.

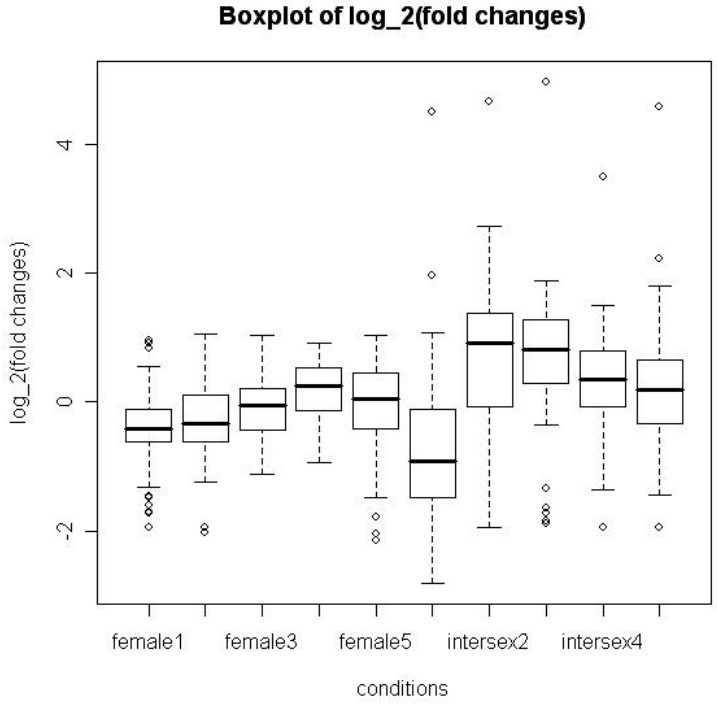

Figure 3.7: Boxplot of Intersex/Female log_2(Fold Change data) by replicate.

It is shown in the boxplot (Figure 3.7) that that there is large variation within each of the intersex and female samples. Also, the outlying dots for the intersex boxplots indicate that there are genes that contribute to great variability when comparing the genes from the two sex groups. Figure 3.8 below will provide a better look at the gene data.

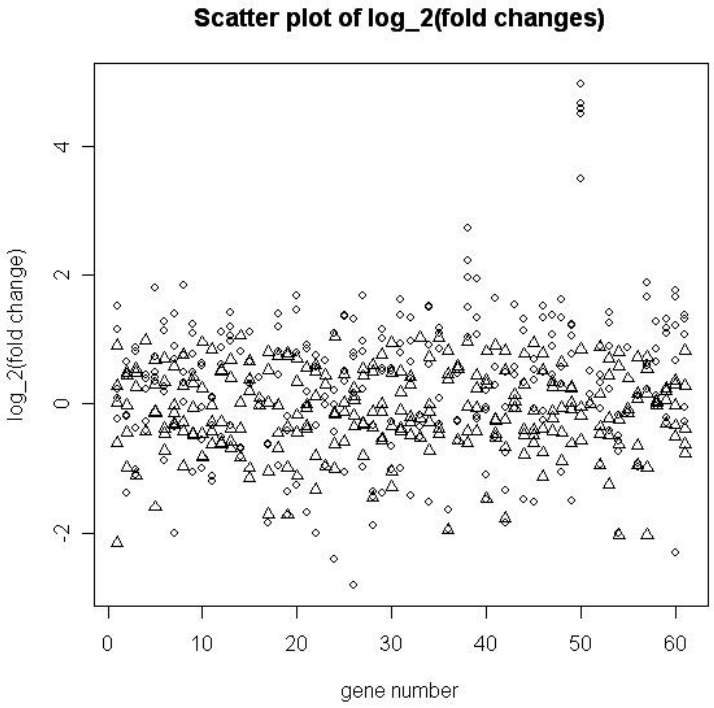

Figure 3.8: Scatter plot of Intersex/Female log_2(Fold Change data) by gene number.

Figure 3.8 shows the fold changes of each gene from the ten replicates; five replicates from each sex group. The triangles represent the fold changes from the female group, while the circles represent the ones from the intersex group. From this graph, we see that the gene which contributed to the most variability is the gene numbered 50. There are other genes that acted quite differently between the female and the intersex groups.

This data was used to examine the three algorithms, LASSO, SLR, and SMLR. The data was processed by each of the three algorithms to determine whether they selected the genes that are responsible for the most variation.

The first algorithm tested was LASSO. After trying out different tuning parameters, 0.8 was chosen. With this tuning parameter, nine genes were selected. Figure 3.9 shows the genes selected in red.

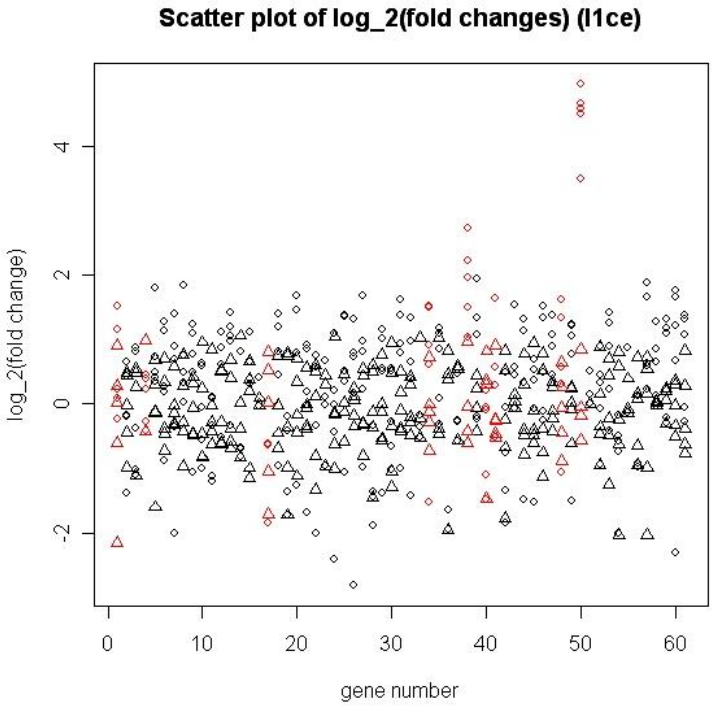

Figure 3.9: Scatter plot of Intersex/Female log_2(Fold Change data) with genes selected by LASSO.

From the plot, we see that the gene numbered 50 was selected. This is the gene displayed in Figure 3.7 with large variability. Some other genes with larger fold changes were also chosen by LASSO. A closer look of the selected genes is provided in Table 3.6 below.

Gene #	Gene name	Female1	Female2	Female3	Female4	Female5	Median
1	FI6	1.22	0.66	1.02	1.88	0.22	1.02
4	FJ35	0.75	0.75	2.00	0.75	0.75	0.75
17	FM34	0.31	1.76	1.43	1.01	0.49	1.01
34	FH22	0.83	0.61	1.00	1.65	0.92	0.92
38	FM68	0.66	0.66	1.97	0.75	0.97	0.75
40	FN44	0.36	1.76	1.29	1.23	0.36	1.23
41	FH2	1.89	0.73	0.70	0.83	0.85	0.83
48	FN26	0.96	0.54	0.74	1.57	1.19	0.96
50	FM40	0.67	1.80	0.67	0.88	0.97	0.88

Gene #	Gene name	Intersex1	Intersex2	Intersex3	Intersex4	Intersex5	Median
1	FI6	0.85	2.85	1.06	2.22	1.18	1.18
4	FJ35	0.83	1.37	1.31	0.75	1.17	1.17
17	FM34	0.28	0.28	0.28	0.64	0.66	0.28
34	FH22	0.35	1.89	1.52	2.81	2.85	1.89
38	FM68	3.88	6.63	2.85	2.05	4.70	3.88
40	FN44	0.36	0.47	0.96	1.16	0.94	0.94
41	FH2	0.70	1.49	1.22	0.70	3.12	1.22
48	FN26	1.25	2.52	3.07	0.48	1.49	1.49
50	FM40	22.78	25.18	31.31	11.28	23.96	23.96

Table 3.6: Table with the fold changes of a subset of the Intersex/Female data that corresponds to the genes selected by LASSO.

According to the median of fold changes given, only one of the nine selected genes are down-regulated genes, while two are up-regulated. This supports the idea suggested in Section 3.1 that LASSO may be more sensitive to up-regulated genes. The fold changes were then compared to their corresponding model coefficients calculated by LASSO given in Table 3.7 below. Again, there seems to be no obvious relation between the magnitude of the fold changes and their corresponding coefficients.

Gene #	Gene name	Coef
	Intercept	0.313
1	FI6	0.060
4	FJ35	-0.026
17	FM34	-0.111
34	FH22	0.133
38	FM68	0.000
40	FN44	-0.046
41	FH2	-0.112
48	FN26	-0.259
50	FM40	0.047

Table 3.7: Table with the coefficients of the regression model computed by LASSO using the Intersex/Female data.

The next algorithm that was tested using this pair of data was SLR. After trying out different tuning parameter values, 0.5 was chosen, and only one gene was selected by SLR. This is the gene numbered 50, which was the one displayed with large variability in Figure 3.7. Again, with only one gene chosen, it is impossible to comment on its sensitivity to up-regulated or down-regulated genes. However, this supports the statement made in Section 3.1 that this algorithm is not very sensitive to its tuning parameter when there is a gene that is obviously different from the others occurs in the data. Various tuning parameters were tried when processing the data with SLR, but the results remained identical with only one gene selected. This algorithm only selected the gene that contributed to great variation, while neglecting the others that might be meaningful but varied less comparatively. Figure 3.10 below shows genes chosen by SLR in red.

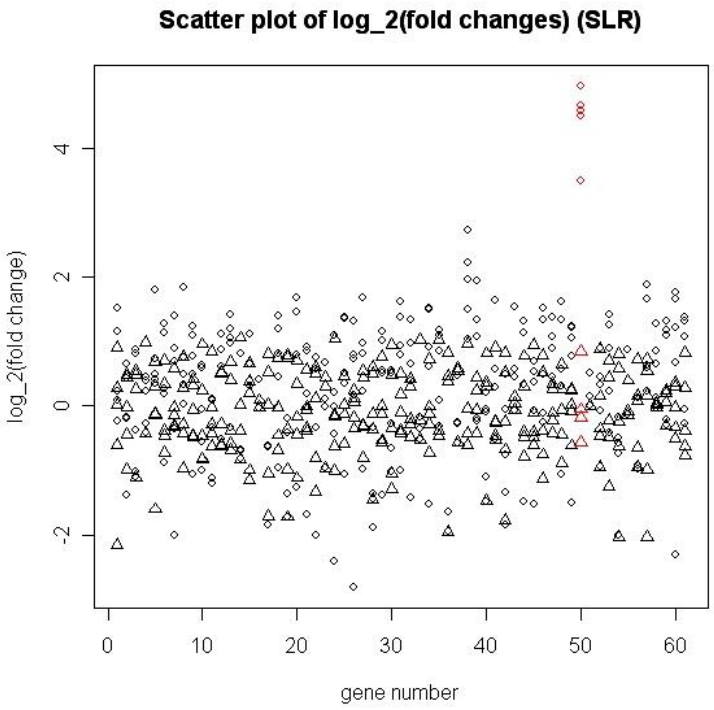

Figure 3.10: Scatter plot of Intersex/Female log_2(Fold Change data) with genes selected by SLR.

Finally, this pair of data was processed with the algorithm SMLR. While setting the tuning parameter to 0.32, 10 genes were selected by the algorithm. Figure 3.11 below shows the genes chosen by the algorithm in red.

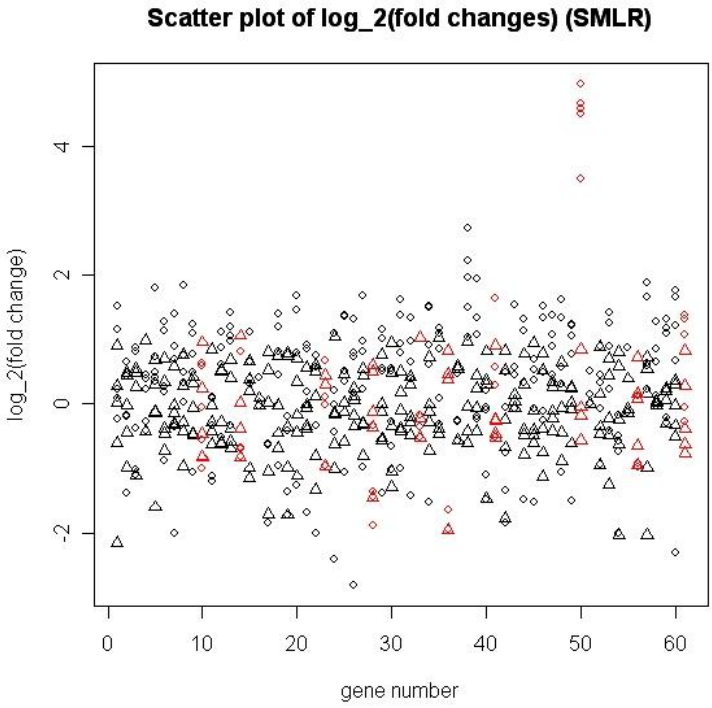

Figure 3.11: Scatter plot of Intersex/Female log_2(Fold Change data) with genes selected by SMLR.

Similar to the other two algorithms, gene number 50 was again selected by SMLR. The set of genes chosen by this algorithm is different from the ones chosen by LASSO. Figure 3.11 above showed that this algorithm selected both up-regulated and down-regulated genes. Table 3.8 below provides a detailed look of the genes selected by SMLR.

Gene #	Gene name	Female1	Female2	Female3	Female4	Female5	Median
10	FJ2	1.95	0.57	0.57	0.73	1.19	0.73
14	FJ23	1.02	2.08	0.57	0.77	0.57	0.77
23	FC5	0.51	1.24	0.51	1.37	1.37	1.24
28	FD3	0.37	1.42	1.51	0.92	0.78	0.92
33	FD43	0.70	2.04	0.86	0.70	0.70	0.70
36	FJ8	0.26	0.26	1.39	1.78	1.31	1.31
41	FH2	1.89	0.73	0.70	0.83	0.85	0.83
50	FM40	0.67	1.80	0.67	0.88	0.97	0.88
56	FN19	1.07	1.65	1.12	0.52	0.64	1.07
61	FG19	1.77	0.65	0.59	0.77	1.22	0.77

Gene #	Gene name	Intersex1	Intersex2	Intersex3	Intersex4	Intersex5	Median
10	FJ2	0.97	1.51	1.55	0.50	0.68	0.97
14	FJ23	0.63	0.57	1.75	0.62	0.57	0.62
23	FC5	0.51	1.07	1.59	0.51	0.99	0.99
28	FD3	0.27	0.80	0.27	0.39	0.37	0.37
33	FD43	0.70	0.70	0.89	0.78	0.70	0.70
36	FJ8	0.26	0.26	0.32	0.26	0.26	0.26
41	FH2	0.70	1.49	1.22	0.70	3.12	1.22
50	FM40	22.78	25.18	31.31	11.28	23.96	23.96
56	FN19	0.53	0.51	1.10	0.51	0.51	0.51
61	FG19	2.09	2.61	2.48	0.97	0.82	2.09

Table 3.8: Table with the fold changes of a subset of the Intersex/Female data that corresponds to the genes selected by SMLR.

Table 3.8 shows that among the ten genes selected by SMLR, only two of them are up-regulated genes, which are the ones obviously shown in Figure 3.11. Another two of them are down-regulated, while the rest are non-regulated genes. This result shows that this algorithm is not more sensitive to either up-regulated or down-regulated genes.

Table 3.9 below provides the coefficients of the multinomial logistic regression model computed using SMLR. When comparing the coefficients with the fold changes given in Table 3.8, we discovered a pattern as to how SMLR determined coefficients. SMLR defined coefficients that correspond to up-regulated genes as negative numbers, and the opposite for the down-regulated genes. The magnitude of the coefficient for each gene was also related to the fold change of the gene. The magnitude of the coefficients were larger when the gene varied more between samples from the two sex groups.

Gene #	Gene name	Coef
10	FJ2	0.551
14	FJ23	0.302
23	FC5	0.004
28	FD3	0.419
33	FD43	0.077
36	FJ8	0.826
41	FH2	0.125
50	FM40	-0.424
56	FN19	0.731
61	FG19	0.035

Table 3.9: Table with the coefficients of the regression model computed by SMLR using the Intersex/Female data.

This is more easily visualized in Figure 3.12 below. In this figure, fold changes of the genes selected by SMLR were plotted against their corresponding coefficients. Of the two up-regulated genes, one has a negative coefficient, while the other one has a positive coefficient. However, the up-regulated gene with positive coefficient has a median fold change of 2.09, which is very close to being a non-regulated gene (0.5<fold change<2). Also, its coefficient is very close to zero. This may explain why this up-regulated gene has a positive coefficient. When looking at the down-regulated genes, their coefficients are both positive. Also, whenever the fold change of a down-regulated gene is smaller, the magnitude of its corresponding coefficient is larger.

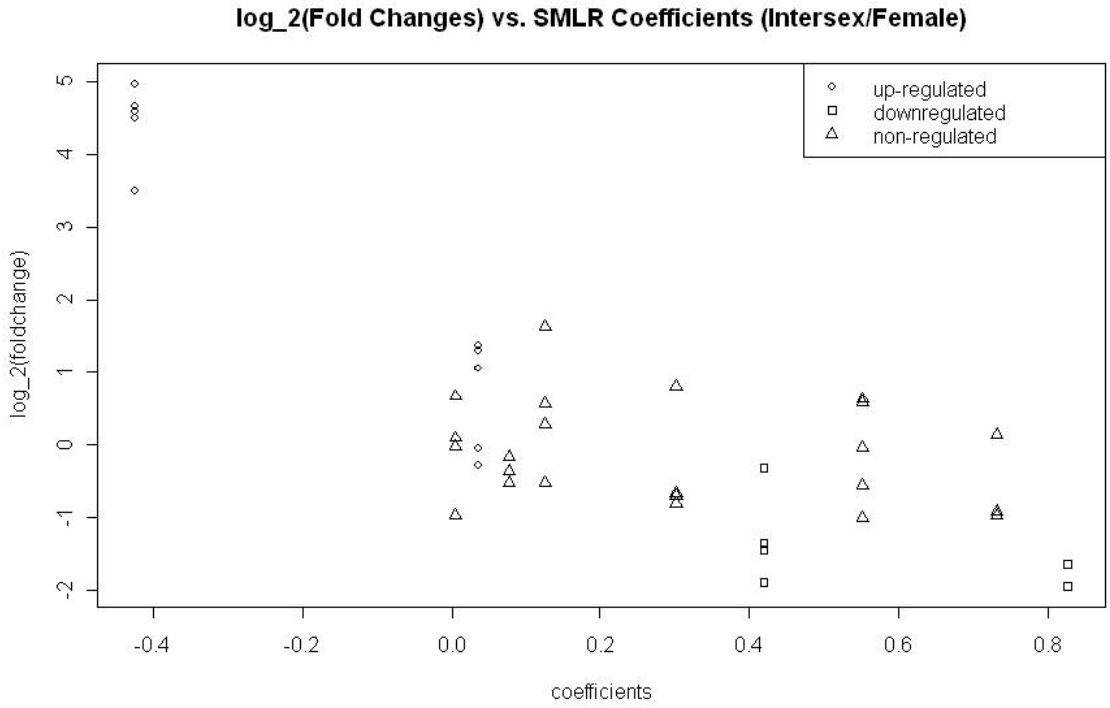

Figure 3.12: Scatter plot of Intersex/Female log_2(Fold change data) of selected genes against their corresponding SMLR model coefficients.

Algorithm	gene #																	
LASSO	1	4			17				34		38	40	41	48	50			
SLR															50			
SMLR			10	14		23	28	33		36			41		50	56	61	

Table 3.10: Table with the number of all the genes selected by each algorithm using the Intersex/Female data.

Table 3.10 shows the number of all the genes selected by each algorithm. There is only one gene that was selected by all of the three algorithms, while there are two of them that were chosen by both LASSO and SMLR.

3.3 Intersex/Male Experiment

The last pair of the Intersex data that was used to examine the three algorithms was the male and intersex pair. The fold changes were computed by dividing the replicates from both sex

groups by the average of the replicates from the male group. Figure 3.13 is given to display an overview of the data. The base-2 logarithm transformed fold changes of all genes from each replicate are included in this boxplot.

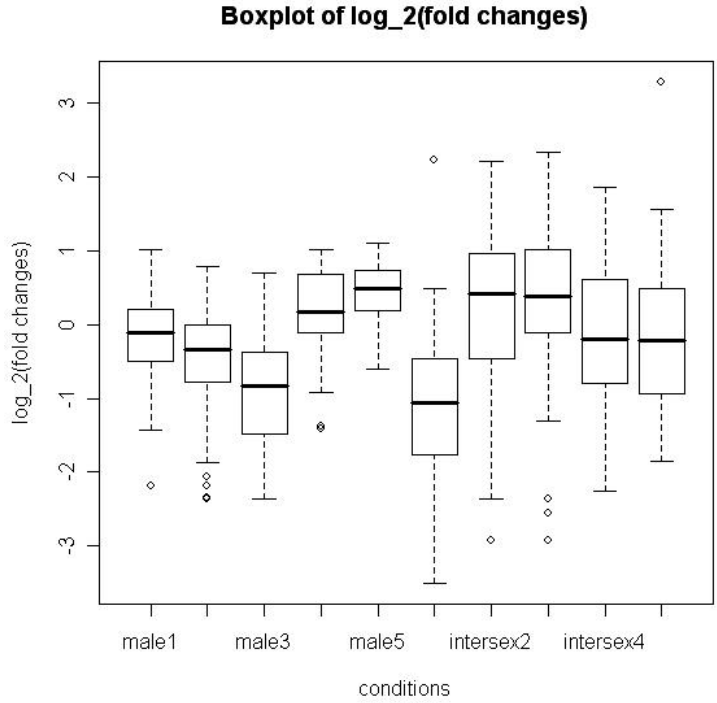

Figure 3.13: Boxplot of Intersex/Male log_2(Fold Change data) by replicate.

From Figure 3.13, we see that although there are large variations within each of the male and intersex samples, the outlying dots for the intersex samples indicate that there is at least one gene that acted very differently between the two sex groups. Other than that, there are a few genes that varied comparatively less, but may also contribute significantly to the variability. Figure 3.14 provides a better look at the gene data.

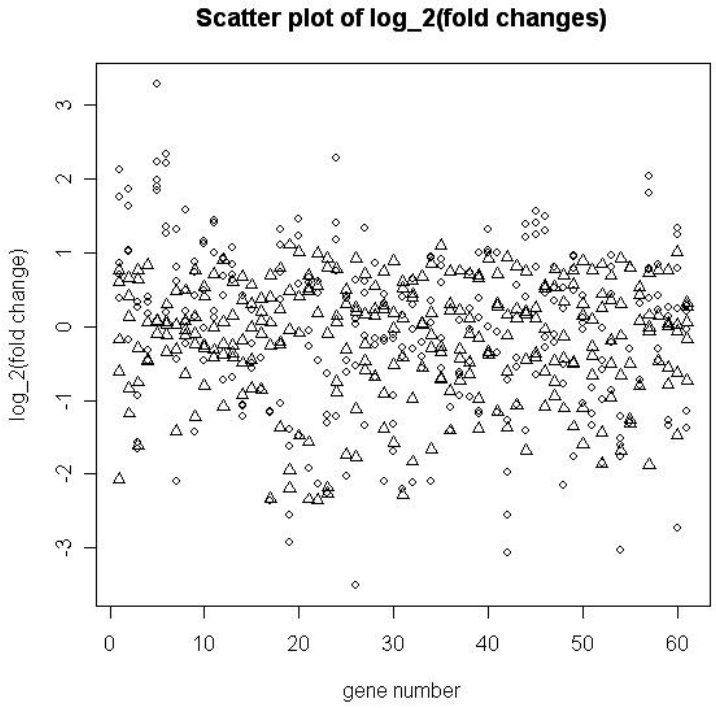

Figure 3.14: Scatter plot of Intersex/Male log_2(Fold Change data) by gene number.

Figure 3.14 shows the fold changes of each gene from all the replicates from both sex groups. The triangles represent the fold changes from the male group, while the circles represent the ones from the intersex group. From this graph, we see that the gene which contributed to the most variability is gene number 5. There are also other genes that acted quite differently between the male and the intersex group.

This data was used to examine the three algorithms, LASSO, SLR, and SMLR. The data was processed by each of the three algorithms to test their ability in selecting genes that were responsible for the most variation.

The first algorithm tested was LASSO. After trying out different tuning parameter values, 0.8 was chosen. With this tuning parameter, eight genes were selected. Figure 3.15 shows the genes selected in red.

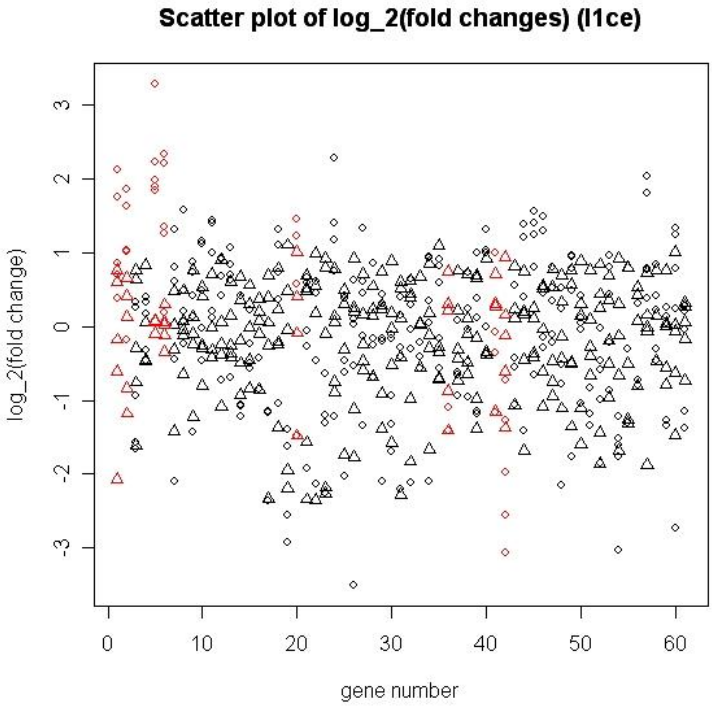

Figure 3.15: Scatter plot of Intersex/Male log_2(Fold Change data) with genes selected by LASSO.

It is shown in Figure 3.15 that the gene numbered 5 was chosen by LASSO. It is the gene that was displayed with large variability in Figure 3.13. In addition, some other genes with larger fold changes as we saw from the Figure 3.13 and Figure 3.14 were selected. A closer look of the data is provided in Table 3.11 showing the fold changes of the genes selected by LASSO.

Gene #	Gene name	Male1	Male2	Male3	Male4	Male5	Median
1	FI6	0.88	0.24	1.52	1.69	0.66	0.88
2	FJ7	0.44	1.09	1.33	0.56	1.57	1.09
5	FJ36	0.93	1.04	1.06	0.93	1.04	1.04
6	FJ39	1.23	0.79	0.92	1.04	1.02	1.02
20	FA10	2.02	0.36	0.36	0.94	1.32	0.94
36	FJ8	0.54	1.68	1.24	0.38	1.16	1.16
41	FH2	1.65	0.45	0.45	1.21	1.24	1.21
42	FI11	1.12	0.92	0.39	1.92	0.66	0.92

Gene #	Gene name	Intersex1	Intersex2	Intersex3	Intersex4	Intersex5	Median
1	FI6	1.30	4.37	1.62	3.40	1.81	1.81
2	FJ7	0.88	3.09	2.02	3.64	2.05	2.05
5	FJ36	4.69	3.96	3.75	3.59	9.81	3.96
6	FJ39	1.14	4.64	5.06	2.55	2.39	2.55
20	FA10	0.36	2.76	2.34	1.49	0.36	1.49
36	FJ8	0.38	0.38	0.47	0.38	0.38	0.38
41	FH2	0.45	0.95	0.78	0.45	2.00	0.78
42	FI11	0.12	0.25	0.17	0.41	0.60	0.25

Table 3.11: Table with the fold changes of a subset of the Intersex/Male data that corresponds to the genes selected by LASSO.

Among the eight genes selected by LASSO, three of them are up-regulated, and two are down-regulated genes. This once again allows us to reasonably suspect that this algorithm is more sensitive to up-regulated genes. The fold changes were then compared to their corresponding model coefficients calculated by LASSO given in the Table 3.12 below. Again, there seems to be no obvious relationship between the magnitude of the fold changes and their corresponding coefficients.

Gene #	Gene name	Coef
	Intercept	0.284
1	FI6	0.006
2	FJ7	0.019
5	FJ36	0.129
6	FJ39	0.071
20	FA10	0.057
36	FJ8	-0.189
41	FH2	-0.299
42	FI11	-0.031

Table 3.12: Table with the coefficients of the regression model computed by LASSO using the Intersex/Male data.

This pair of data was then employed to test SLR. Only one gene was selected by SLR with the tuning parameter 0.5. The gene selected is gene numbered 5, which is the gene that acted the most differently between individuals from the two sex groups. Again, it is impossible to comment on its sensitivity to up-regulated or down-regulated genes when only one gene was selected by the algorithm. However, this again provides evidence to suspect that the algorithm is not very sensitive to its tuning parameter when there is a gene that is obviously different from the others occurs in the data. When the data was processed with SLR, various tuning parameters were tried, but the results remained the same with only one gene selected. It seems that the algorithm only selected the gene that contributed to great variation, while neglecting the others that might be meaningful but with comparatively less variation. Figure 3.16 shows genes selected by SLR in red.

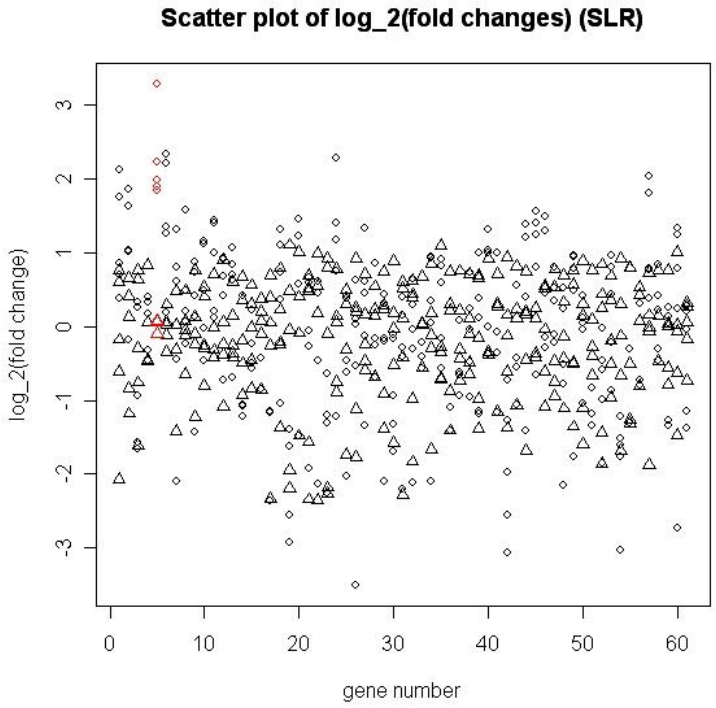

Figure 3.16: Scatter plot of Intersex/Male log_2(Fold Change data) with genes selected by SLR.

Finally, this pair of data was processed with the algorithm SMLR. While setting the tuning parameter to 0.32, 7 genes were chosen by the algorithm. The scatter plot below (Figure 3.17) shows the genes selected by the algorithm in red.

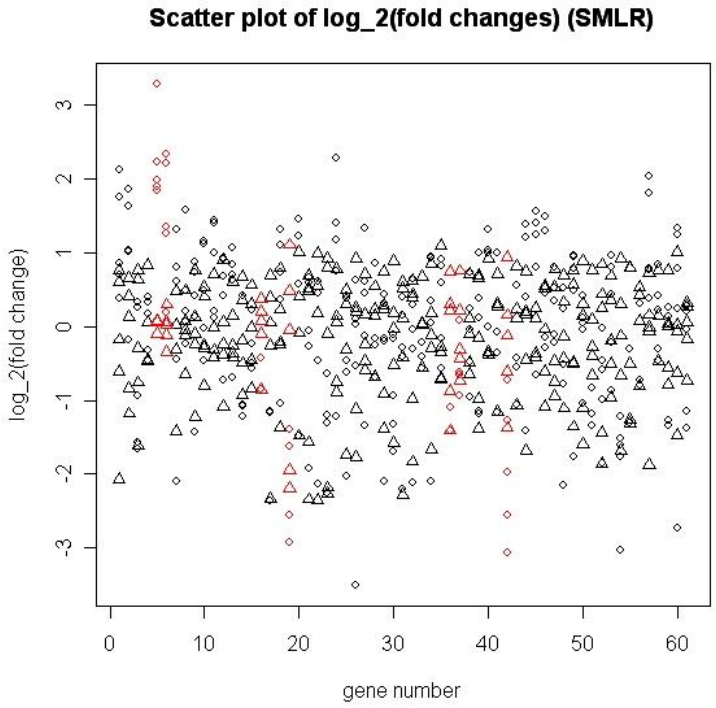

Figure 3.17: Scatter plot of Intersex/Male log_2(Fold Change data) with genes selected by SMLR.

It is not a surprise to see that gene number 5 was chosen. Again, the set of genes selected by SMLR algorithm is different from the one selected by LASSO. We see from Figure 3.17 that among the seven genes selected by SMLR, there are both up-regulated and down-regulated genes. Table 3.13 provides a better look at the gene data.

Gene #	Gene name	Male1	Male2	Male3	Male4	Male5	Median
5	FJ36	0.93	1.04	1.06	0.93	1.04	1.04
6	FJ39	1.23	0.79	0.92	1.04	1.02	1.02
16	FK25	1.06	0.56	1.15	0.93	1.30	1.06
19	FM24	0.97	0.22	0.26	1.39	2.16	0.97
36	FJ8	0.54	1.68	1.24	0.38	1.16	1.16
37	FM61	1.69	0.60	0.81	1.16	0.74	0.81
42	FI11	1.12	0.92	0.39	1.92	0.66	0.92

Gene #	Gene name	Intersex1	Intersex2	Intersex3	Intersex4	Intersex5	Median
5	FJ36	4.69	3.96	3.75	3.59	9.81	3.96
6	FJ39	1.14	4.64	5.06	2.55	2.39	2.55
16	FK25	0.56	0.56	0.56	0.56	0.75	0.56
19	FM24	0.17	0.13	0.13	0.38	0.33	0.17
36	FJ8	0.38	0.38	0.47	0.38	0.38	0.38
37	FM61	0.52	0.52	1.06	0.65	0.64	0.64
42	FI11	0.12	0.25	0.17	0.41	0.6	0.25

Table 3.13: Table with the fold changes of a subset of the Intersex/Male data that corresponds to the genes selected by SMLR.

Table 3.13 shows that among the seven genes selected by SMLR, two of them are up-regulated genes, and three are down-regulated. This time, SMLR selected more down-regulated genes than up-regulated genes.

Table 3.14 below provides the coefficients of the multinomial logistic regression model computed using SMLR. When comparing the coefficients with the fold changes given in Table 3.13, it shows that SMLR might define coefficients in a regular pattern. It seems that SMLR determined coefficients that correspond to up-regulated genes are negative, while opposite holds for the down-regulated genes. The magnitude of the coefficient for each gene was also related to the fold change of the gene. The magnitude of the coefficient is larger when the gene varied more between the two sex groups.

Gene #	Gene name	Coef
5	FJ36	-1.116
6	FJ39	-0.031
16	FK25	0.322
19	FM24	0.374
36	FJ8	1.685
37	FM61	0.214
42	FI11	1.068

Table 3.14: Table with the coefficients of the regression model computed by SMLR using the Intersex/Male data.

Figure 3.18 below provides a better view of this relationship. It can be seen on the left of the plot that for genes with log_2(fold changes) greater than one (up-regulated), their corresponding coefficients are negative. Also, it is shown in the figure that when the log_2(fold change) of a down-regulated gene is smaller, its corresponding coefficient is larger. Moreover, the coefficients of the non-regulated genes are closer to zero.

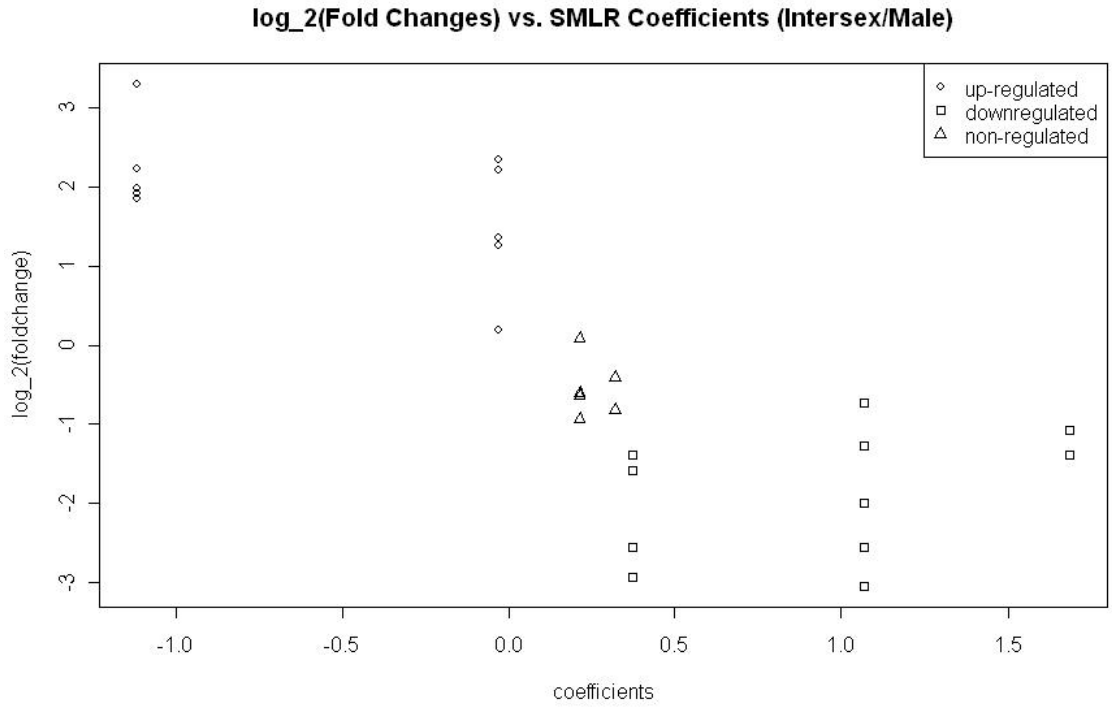

Figure 3.18: Scatter plot of Intersex/Male log_2(Fold change data) of selected genes against their corresponding SMLR model coefficients.

Algorithm	gene #										
LASSO	1	2	5	6			20	36		41	42
SLR			5								
SMLR			5	6	16	19		36	37		42

Table 3.15: Table with the number of all the genes selected by each algorithm using the Intersex/Male data.

Table 3.15 shows the number of all the genes chosen by each algorithm. There is only one gene that were selected by all of the three algorithms, while four of them were selected by both LASSO and SMLR.

3.4 Male/Female/Intersex Experiment

After comparing the behaviour of the three algorithms with the different pairs of the Intersex data, the complete Intersex data with all of the three categories (Male, Female, and Intersex) were processed. However, because SMLR is the only algorithm that can accommodate data with more than two categories, this experiment focused on its ability to handle this type of data.

Experiment	gene #													
Male/Female				5		10	11				19			24	28
Female/Intersex	1		4			10		14		17			23		28
Male/Intersex	1	2		5	6				16		19	20			
Male/Female/Intersex	1			5	6	10	11	14	16		19		23		28

Experiment	gene #													
Male/Female						38	39	40		42	43		49	50	
Female/Intersex	33	34	36			38		40	41			48		50	
Male/Intersex			36	37					41	42					
Male/Female/Intersex			36			38	39		41	42				50	51

Experiment	gene #				
Male/Female	54	56	57		
Female/Intersex		56			61
Male/Intersex					
Male/Female/Intersex		56	57	60	61

Table 3.16: Table with a summary of the number of all the genes selected by all algorithms in each experiment

Table 3.16 provides a summary of the genes selected by all of the three algorithms in each of the experiments performed in this chapter. The gene number of the selected genes are given. The bottom row contains the numbers of the genes selected by SMLR in this section using the complete data with all of the three sex groups. When comparing this row to the other three rows, it is not difficult to realize that most of the genes selected by the three algorithms in the previous sections in this chapter using different pairs of data were covered by SMLR in this three categories experiment. This may suggest that SMLR has the ability to handle data with more than two categories. It summarized the results gathered from the other experiments carried out in this chapter that processed these three categories in pairs using all the three algorithms. A further investigation of SMLR in handing data with more than two categories will be provided in following chapters.

4 Breast Cancer Data Experiment

The second dataset that was used to compare the results from the three algorithms is the Breast Cancer data. This data was downloaded from the website of Duke Institute for Genome Sciences & Policy (2007). West et al. (2001) explained in their article that primary breast tumors were obtained from the Duke Breast Cancer SPORE frozen tissue bank. Tumors were either positive for both the estrogen and progesterone receptors or negative for both receptors. All tumors were diagnosed as invasive ductal carcinoma and were between 1.5 and 5 cm in maximal dimension. In each case, a diagnostic axillary lymph node dissection was performed. Each potential tumor was examined by hematoxylinyeosin staining and only those that were >60% tumor (on a per-cell basis), with few infiltrating lymphocytes or necrotic tissue, were carried on for RNA extraction. The final collection of tumors consisted of 13 estrogen receptor positive(ER+) lymph node positive(LN+) tumors, 12 ER negative LN+ tumors, 12 ER+LN- tumors, and 12 ER-LN- tumors.

Since LASSO, and SLR can only accommodate data with two categories, the four conditions of the Breast Cancer data (ER+LN+, ER+LN-, ER-LN+, ER-LN-) were separated into two pairs (ER+/ER-, LN-/LN+). Moreover, because of the background noise, there were negative intensity values in the data. The data was shifted up by adding the minimum plus 50 to the data to ensure that all values were greater than zero. After shifting, the data was transformed from intensities into fold changes. The fold changes were computed by dividing the intensities of each replicate of both groups by the average of intensities of all replicates of one of the two groups. Furthermore, because this data includes 7129 genes, and one of the R algorithms, the LASSO routine can only handle vectors up to length 5000, a random sample of 5000 genes was selected from the data. After all the modifications, each pair of data was then used to test and compare the ability of the three different algorithms in selecting feature genes. The complete Breast Cancer dataset was also employed to examine SMLR's ability to work with data that has more than two categories.

4.1 ER+/ER- Experiment

The first pair of the Breast Cancer data that was examined using the three algorithms is the ER+ and ER- pair. The fold changes were computed by dividing the replicates from both groups by the average of the replicates from the ER- group. Figure 4.1 provides an overview of the data by plotting the base-2 logarithm transformed fold changes.

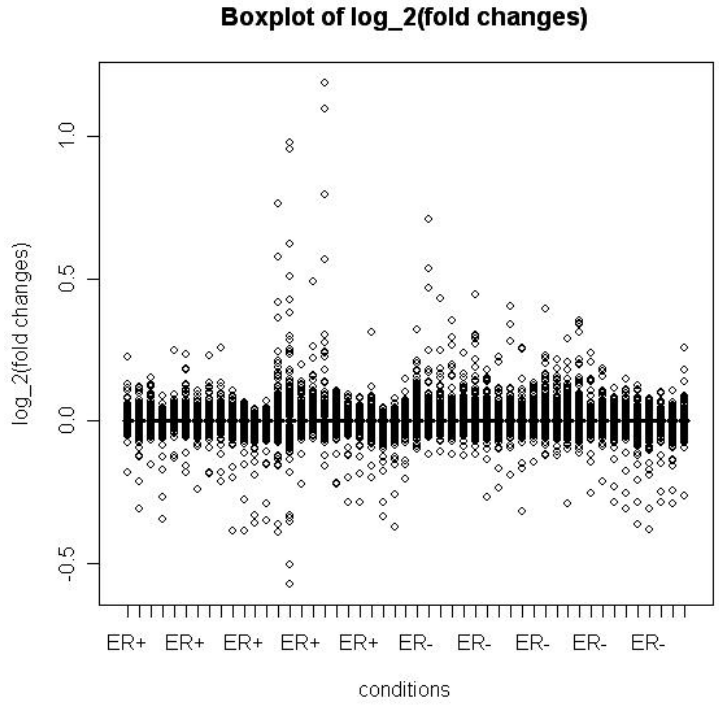

Figure 4.1: Boxplot of ER+/ER- log_2(Fold Change data) by replicate.

From Figure 4.1, we see that there are slight differences between the genes from the ER+ and the ER- tumor. However, in contrast to the Intersex data, where the fold changes range from 0.1 to 24, the fold changes of this data are mostly within the range 0.5 to 2, which is defined in the previous chapter as the range of non-regulated genes. If the same definition is used in this chapter, probably none of the genes will be identified as regulated genes. Therefore, in this chapter, genes with median fold change above 1 will be treated as up-regulated genes, and the genes with median fold changes below 1 are going to be treated as down-regulated genes. Figure 4.2 provides a better look at this gene data.

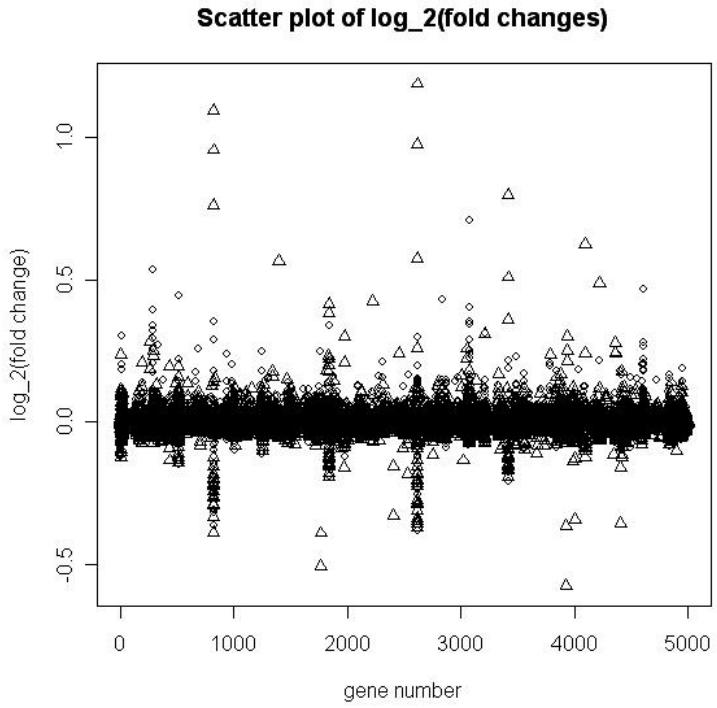

Figure 4.2: Scatter plot of ER+/ER- log_2(Fold Change data) by gene number.

Figure 4.2 shows the base-2 logarithm transformed fold changes of each gene from the 49 replicates; 25 from the ER+ group and 24 from the ER- group. The triangles represent the fold changes from the ER+ group, while the circles represent the ones from the ER- group. From this graph, we see that there are a few genes that may be regulated.

This data was then employed to examine the three algorithms. The data was processed by each of the three algorithms to access their ability in selecting the genes that were responsible for the most variation.

The first algorithm tested was LASSO. With only a range of tuning parameter values that allowed the algorithm to stay converged, the one that allowed the algorithm to select the most genes was employed. The tuning parameter, 1.15 was used in this experiment. With this tuning parameter, three genes were chosen. Figure 4.3 shows the selected genes in red.

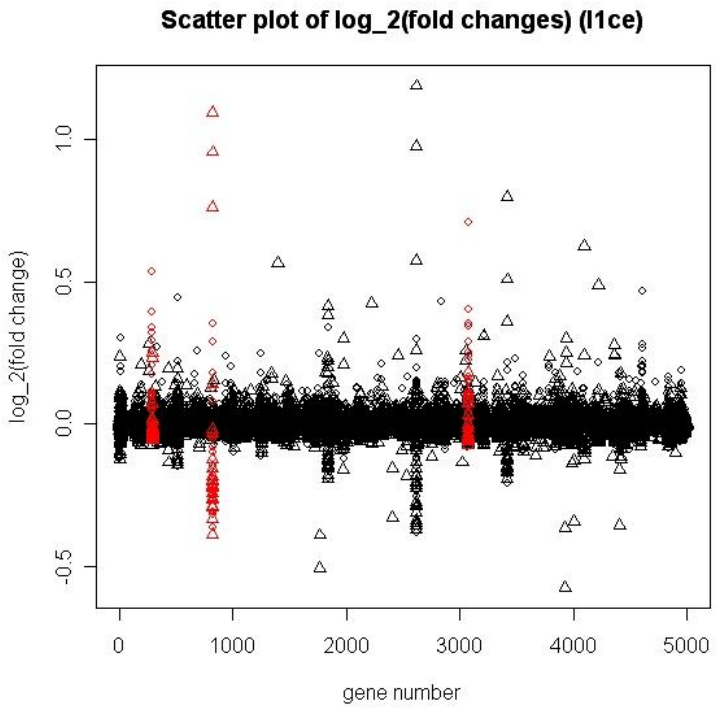

Figure 4.3: Scatter plot of ER+/ER- log_2(Fold Change data) with genes selected by LASSO.

It is shown in Figure 4.3 that only one of the three genes is down-regulated, while the other two are up-regulated. Also, there were a few other genes that seemed to vary more were not selected. We may suspect that LASSO is not very accurate at choosing differentially expressed genes when the overall gene set does not change too much over conditions. A closer look of the data is provided in Table 4.1 and Table 4.2 below. This supports the idea given in the previous chapter that LASSO is more sensitive to up-regulated genes.

Gene #	Gene name	ER+ #1	ER+ #2	ER+ #3	ER+ #4	ER+ #5
292	M87789_s_at	0.98	0.96	0.96	0.97	1.19
823	AFFX-CreX-5_at	0.98	0.86	0.99	0.83	0.92
3075	M34516_at	1.05	0.96	0.97	0.97	1.07

Gene #	Gene name	ER+ #6	ER+ #7	ER+ #8	ER+ #9	ER+ #10
292	M87789_s_at	1.05	0.98	1.17	0.99	0.99
823	AFFX-CreX-5_at	0.90	0.85	0.88	0.88	0.77
3075	M34516_at	1.13	0.96	1.07	0.98	0.96

Gene #	Gene name	ER+ #11	ER+ #12	ER+ #13	ER+ #14	ER+ #15
292	M87789_s_at	0.97	1.01	0.98	0.98	1.01
823	AFFX-CreX-5_at	0.76	0.79	0.82	1.70	1.94
3075	M34516_at	0.97	0.97	0.99	0.95	1.00

Gene #	Gene name	ER+ #16	ER+ #17	ER+ #18	ER+ #19	ER+ #20
292	M87789_s_at	0.98	0.98	0.97	0.97	1.01
823	AFFX-CreX-5_at	0.86	1.10	2.14	0.86	0.87
3075	M34516_at	1.08	1.00	1.05	0.97	1.02

Gene #	Gene name	ER+ #21	ER+ #22	ER+ #23	ER+ #24	ER+ #25
292	M87789_s_at	0.98	0.99	0.96	0.99	0.96
823	AFFX-CreX-5_at	0.88	0.87	0.82	0.84	0.90
3075	M34516_at	0.97	0.97	0.97	0.99	0.97

Gene #	Gene name	Median
292	M87789_s_at	0.98
823	AFFX-CreX-5_at	0.87
3075	M34516_at	0.97

Table 4.1: Table with the fold changes of a subset of the ER+/ER- data that corresponds to the genes selected by LASSO. (ER+ only)

Gene #	Gene name	ER- #1	ER- #2	ER- #3	ER- #4	ER- #5
292	M87789_s_at	1.25	1.45	0.97	1.22	1.00
823	AFFX-CreX-5_at	0.93	1.13	1.09	0.97	0.97
3075	M34516_at	1.07	1.63	0.97	1.28	1.12

Gene #	Gene name	ER- #6	ER- #7	ER- #8	ER- #9	ER- #10
292	M87789_s_at	0.97	1.04	1.07	1.27	1.20
823	AFFX-CreX-5_at	1.22	0.83	0.89	0.88	0.80
3075	M34516_at	0.97	1.11	1.17	1.32	1.19

Gene #	Gene name	ER- #11	ER- #12	ER- #13	ER- #14	ER- #15
292	M87789_s_at	0.98	1.32	0.97	1.08	1.05
823	AFFX-CreX-5_at	0.90	1.06	0.92	0.95	1.28
3075	M34516_at	0.99	1.06	0.98	1.22	1.27

Gene #	Gene name	ER- #16	ER- #17	ER- #18	ER- #19	ER- #20
292	M87789_s_at	1.16	1.13	1.08	1.06	1.06
823	AFFX-CreX-5_at	0.91	0.94	0.85	0.81	0.78
3075	M34516_at	1.18	1.08	1.03	1.11	1.00

Gene #	Gene name	ER- #21	ER- #22	ER- #23	ER- #24	Median
292	M87789_s_at	0.96	1.07	1.03	0.97	1.06
823	AFFX-CreX-5_at	0.81	0.84	0.84	0.83	0.91
3075	M34516_at	0.97	1.07	1.01	0.98	1.07

Table 4.2: Table with the fold changes of a subset of the ER+/ER- data that corresponds to the genes selected by LASSO. (ER- only)

Table 4.3 below provides the coefficients of the regression model computed by LASSO. After comparing these coefficients and the fold changes of the genes chosen by LASSO given in the previous table, there seems to be no obvious relationship between the magnitude of the fold changes and their corresponding coefficients.

Gene #	Gene name	Coef
	Intercept	-0.580
292	M87789_s_at	0.154
823	AFFX-CreX-5_at	-0.071
3075	M34516_at	0.925

Table 4.3: Table with the coefficients of the regression model computed by LASSO using the ER+/ER- data.

The next algorithm we tested using this data is SLR. Again, there is only a range of tuning

parameter values that allows the algorithm to stay converged. With the tuning parameter 1.1, three genes were chosen by SLR. Figure 4.4 shows the selected genes in red.

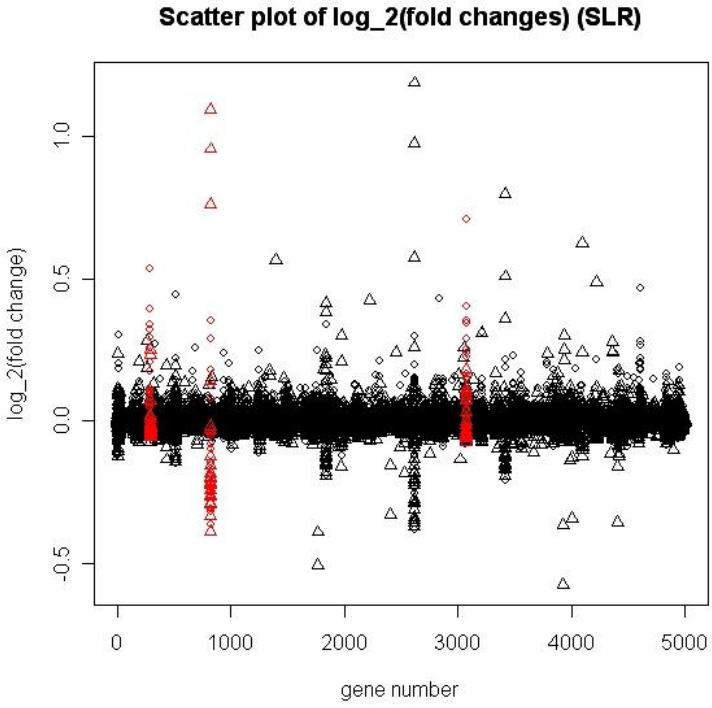

Figure 4.4: Scatter plot of ER+/ER- log_2(Fold Change data) with genes selected by SLR.

The genes selected by SLR are the same as the genes selected by LASSO. It selected one down-regulated gene, and two up-regulated genes. SLR was selecting only one gene in the previous chapter; however, in this experiment, SLR selected more than one gene. This may allow one to think that SLR selects more genes when the data does not change much over conditions. Since SLR selected same genes as LASSO, one may refer back to Table 4.1 and Table 4.2 for a detailed look of the data.

Gene #	Gene name	Coef
	Intercept	-1.106E-01
292	M87789_s_at	6.685E-03
823	AFFX-CreX-5_at	-2.987E-05
3075	M34516_at	6.605E-02

Table 4.4: Table with the coefficients of the regression model computed by SLR using the ER+/ER- data.

Table 4.4 provides the coefficients of the regression model computed by SLR. After comparing these coefficients and the fold changes of the genes chosen by SLR. There seems to be no obvious relationship between the magnitude of the fold changes and their corresponding coefficients.

Finally, the data is processed with the algorithm SMLR. While there is only a range of tuning parameter values that allows the algorithm to stay converged, the tuning parameter 1.2 was use. With this tuning parameter, only one gene was chosen by the algorithm. Figure 4.5 shows the selected gene in red.

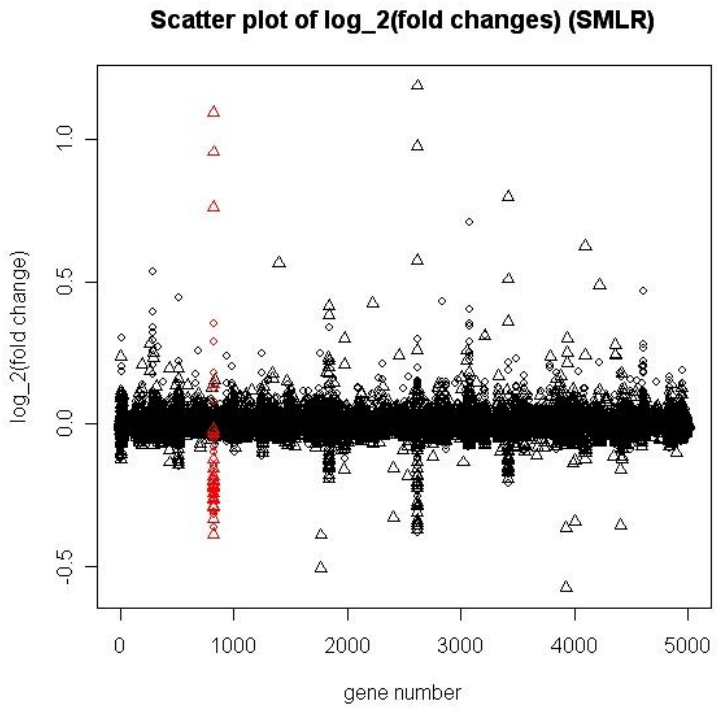

Figure 4.5: Scatter plot of ER+/ER- log_2(Fold Change data) with genes selected by SMLR.

It is not a surprise to see that this gene was selected as this was also selected by LASSO and SLR. Since only one gene was chosen by this algorithm in this experiment, it is impossible to comment on its sensitivity to up-regulated or down-regulated genes. However, it is reasonable to suspect that SMLR only selects a small number of genes when the data does not change much over conditions.

Algorithm	gene #		
LASSO	292	823	3075
SLR	292	823	3075
SMLR		823	

Table 4.5: Table with the number of all the genes selected by each algorithm using the ER+/ER- data.

Table 4.5 provides the number of all the genes selected by each algorithm. There is only one gene that is chosen by all of the three algorithms, while all of them selected by both LASSO and SLR.

4.2 LN-/LN+ Experiment

Another pair of the Breast Cancer data that was examined using the three algorithms is the LN+ and LN- pair. The fold changes were computed by dividing the replicates from both groups by the average of the replicates from the LN+ group. Figure 4.6 provides an overview of the data.

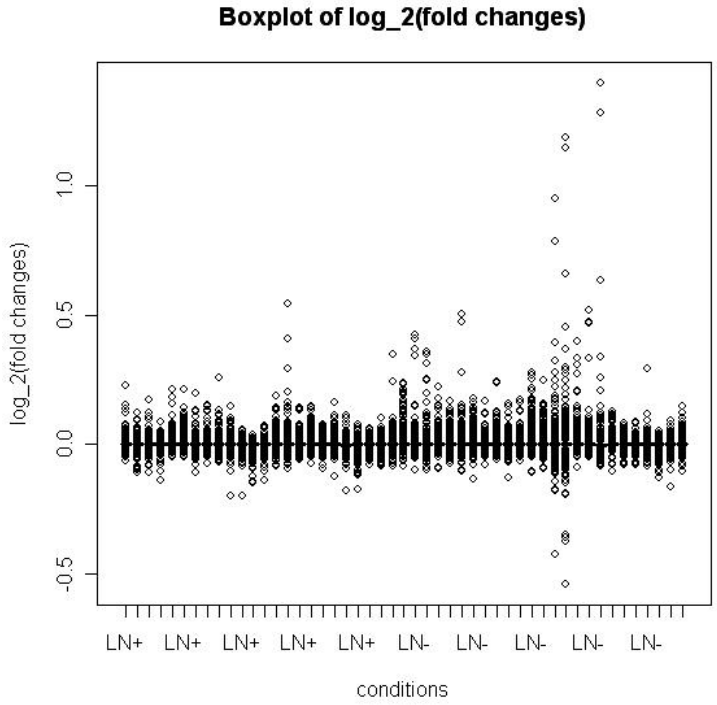

Figure 4.6: Boxplot of LN-/LN+ log_2(Fold Change data) by replicate.

From the boxplot, we see that there are slight differences between the genes from the LN+ and the LN- tumor. Similar to the EN+/EN- data, most of the genes in this data have fold changes close to 1. There are only a few genes that contributed to greater variability when comparing between groups. Figure 4.7 provides a better look at the gene data.

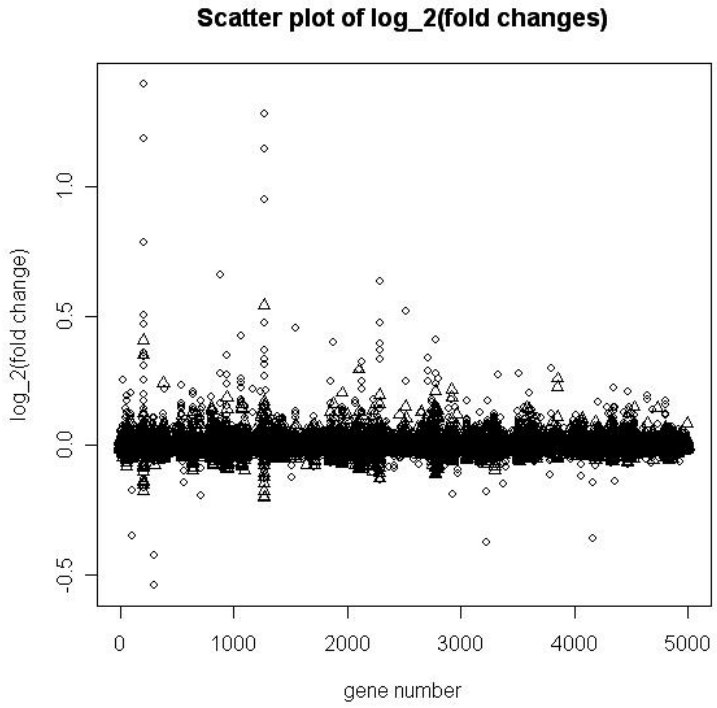

Figure 4.7: Scatter plot of LN-/LN+ log_2(Fold Change data) by gene number.

Figure 4.7 shows the fold changes of each gene from the 49 replicates; 24 replicates from the LN+ group, and 25 replicates from the LN- group. The triangles represent the fold changes from the LN- group, while the circles represent the ones from the LN+ group. From this figure, we see that there are three genes numbered in the range between 1 to 2000 that may contribute to greater variability between conditions.

This data was then used to examine the three algorithms, LASSO, SLR, and SMLR. The data was processed by each of the three algorithms to test their ability to select the genes that are responsible for the most variation between conditions.

The first algorithm tested was LASSO. The tuning parameter, 1.15 was used in this experiment. With this tuning parameter, four genes were chosen. Figure 4.8 shows the selected genes in red.

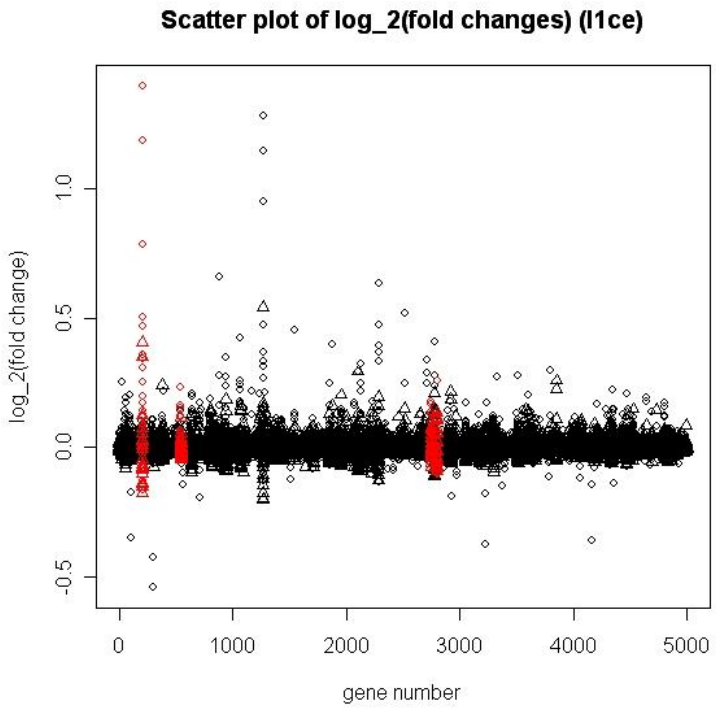

Figure 4.8: Scatter plot of LN-/LN+ log_2(Fold Change data) with genes selected by LASSO.

From Figure 4.8, we can see only one of the three genes mentioned above was chosen, while the three other genes selected do not seem to vary as much. A closer look of the data is provided in Table 4.6 and Table 4.7.

Gene #	Gene name	LN+ #1	LN+ #2	LN+ #3	LN+ #4	LN+ #5
208	hum_alu_at	1.02	0.93	1.04	0.91	1.05
546	AFFX-HUMGAPDH/					
	M33197_M_at	1.00	1.01	0.99	1.02	1.00
2751	HG2873-HT3017_at	1.05	1.00	1.01	1.00	0.99
2802	Z23090_at	1.10	0.99	1.02	1.01	1.05

Gene #	Gene name	LN+ #6	LN+ #7	LN+ #8	LN+ #9	LN+ #10
208	hum_alu_at	1.02	1.09	1.02	1.00	1.01
546	AFFX-HUMGAPDH/					
	M33197_M_at	1.00	0.99	0.97	1.08	0.99
2751	HG2873-HT3017_at	1.03	0.98	0.99	1.05	0.99
2802	Z23090_at	1.08	0.97	0.99	1.08	1.04

Gene #	Gene name	LN+ #11	LN+ #12	LN+ #13	LN+ #14
208	hum_alu_at	0.95	0.90	0.91	0.94
546	AFFX-HUMGAPDH/				
	M33197_M_at	0.98	0.98	0.98	0.99
2751	HG2873-HT3017_at	0.98	0.98	0.97	0.99
2802	Z23090_at	1.01	0.97	0.97	0.95

Gene #	Gene name	LN+ #15	LN+ #16	LN+ #17	LN+ #18
208	hum_alu_at	1.33	0.97	1.00	0.95
546	AFFX-HUMGAPDH/				
	M33197_M_at	1.00	0.99	1.02	1.01
2751	HG2873-HT3017_at	0.96	1.00	1.04	1.05
2802	Z23090_at	0.94	0.97	0.95	0.97

Gene #	Gene name	LN+ #19	LN+ #20	LN+ #21	LN+ #22
208	hum_alu_at	0.97	0.95	0.89	0.95
546	AFFX-HUMGAPDH/				
	M33197_M_at	1.04	0.97	0.98	1.00
2751	HG2873-HT3017_at	1.01	0.95	0.98	1.01
2802	Z23090_at	1.04	0.97	0.95	0.96

Gene #	Gene name	LN+ #23	LN+ #24	Median
208	hum_alu_at	0.94	1.27	0.97
546	AFFX-HUMGAPDH/			
	M33197_M_at	1.01	0.98	1.00
2751	HG2873-HT3017_at	0.97	1.02	1.00
2802	Z23090_at	0.95	1.07	0.98

Table 4.6: Table with the fold changes of a subset of the LN-/LN+ data that corresponds to the genes selected by LASSO (LN+ only).

Gene #	Gene name	LN- #1	LN- #2	LN- #3	LN- #4	LN- #5
208	hum_alu_at	1.15	1.27	1.28	1.07	1.06
546	AFFX-HUMGAPDH/					
	M33197_M_at	1.18	1.03	1.09	1.02	1.08
2751	HG2873-HT3017_at	1.1	1.05	1.01	1.05	1.12
2802	Z23090_at	0.96	0.93	0.94	0.98	0.95

Gene #	Gene name	LN- #6	LN- #7	LN- #8	LN- #9	LN- #10
208	hum_alu_at	1.42	1.11	0.98	1.09	1.04
546	AFFX-HUMGAPDH/					
	M33197_M_at	0.98	1.05	0.99	0.99	1.05
2751	HG2873-HT3017_at	1.09	1.04	1.01	1.07	1.02
2802	Z23090_at	0.93	0.95	0.95	1.00	1.00

Gene #	Gene name	LN- #11	LN- #12	LN- #13	LN- #14	LN- #15
208	hum_alu_at	1.13	1.19	1.09	1.72	2.28
546	AFFX-HUMGAPDH/					
	M33197_M_at	1.12	1.00	1.11	1.01	0.98
2751	HG2873-HT3017_at	1.05	1.13	0.97	1.03	1.00
2802	Z23090_at	1.00	0.93	0.94	0.93	0.94

Gene #	Gene name	LN- #16	LN- #17	LN- #18	LN- #19	LN- #20
208	hum_alu_at	1.24	1.39	2.64	0.99	0.95
546	AFFX-HUMGAPDH/					
	M33197_M_at	1.03	1.00	0.99	1.04	1.03
2751	HG2873-HT3017_at	1.04	1.02	1.02	1.04	1.00
2802	Z23090_at	1.09	0.97	1.20	1.06	0.95

Gene #	Gene name	LN- #21	LN- #22	LN- #23	LN- #24	LN- #25
208	hum_alu_at	0.95	1.08	0.91	0.89	1.00
546	AFFX-HUMGAPDH/					
	M33197_M_at	0.99	1.00	0.97	0.99	1.00
2751	HG2873-HT3017_at	0.97	1.01	0.97	1.01	1.04
2802	Z23090_at	1.00	0.97	0.97	0.97	0.99

Gene #	Gene name	Median
208	hum_alu_at	1.09
546	AFFX-HUMGAPDH/	
	M33197_M_at	1.01
2751	HG2873-HT3017_at	1.03
2802	Z23090_at	0.97

Table 4.7: Table with the fold changes of a subset of the LN-/LN+ data that corresponds to the genes selected by LASSO (LN- only).

Among all the four genes selected by LASSO, three of them are up-regulated (fold change>1), and only one is down-regulated (fold change<1). Again, it provides evidence to suspect that LASSO is more sensitive to up-regulated genes. Also, LASSO only selected one gene that shows an obvious difference, while neglecting some other genes that might possibly be differentially expressed genes. We may suspect that LASSO is not very accurate at choosing differentially expressed genes while the overall gene set does not change too much over conditions. Table 4.8 below provides the coefficients of the regression model computed by LASSO. After comparing these coefficients and the fold changes of the genes chosen by LASSO given in Table 4.6 and Table 4.7, again, there seems to be no obvious relation between the magnitude of the fold changes and their corresponding coefficients.

Gene #	Gene name	Coef
	Intercept	-0.691
208	Hum_alu_at	0.496
546	AFFX-HUMGAPDH/M33197_M_at	0.042
2751	HG2873-HT3017_at	0.602
2802	Z23090_at	-0.010

Table 4.8: Table with the coefficients of the regression model computed by LASSO using the LN-/LN+ data.

The next algorithm we test using this data was SLR. Again, there is only a range of tuning parameter values that allows the algorithm to stay converged. Different tuning parameters were tried, and the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 1.15 was used in this experiment. Only one gene was chosen by SLR. Figure 4.9 shows the selected gene in red.

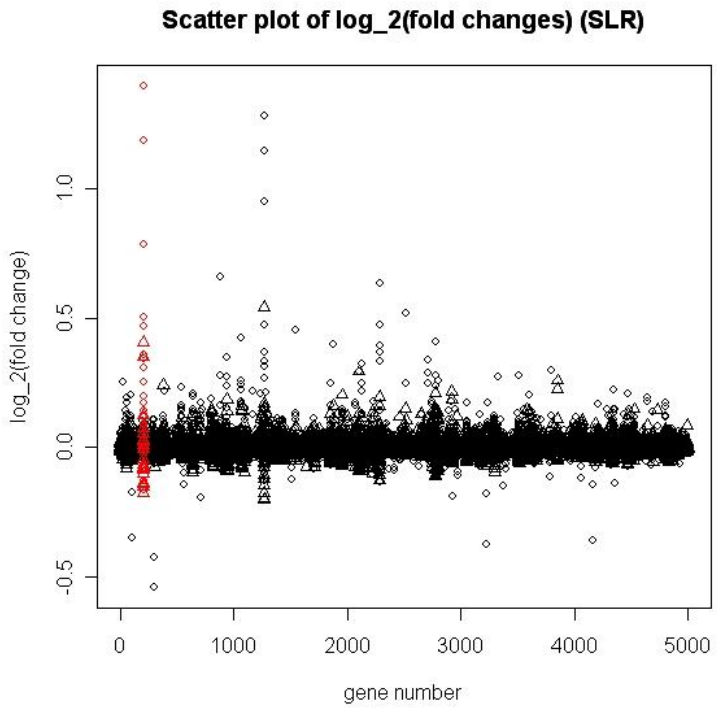

Figure 4.9: Scatter plot of LN-/LN+ log_2(Fold Change data) with genes selected by SLR.

The gene selected by SLR is a member of the list of genes selected by LASSO. This gene has the highest median fold change. Also, it is shown in Figure 4.9 that the gene selected seemed to be one of the genes that has varied the most between groups. Again, it is impossible to comment on its sensitivity to up-regulated or down-regulated genes as only one gene was chosen by this algorithm in this experiment.

Finally, the data was processed with the algorithm SMLR. The tuning parameter 2.0 was used in this experiment. The algorithm selected the same gene as SLR. Figure 4.10 below shows the genes selected in red.

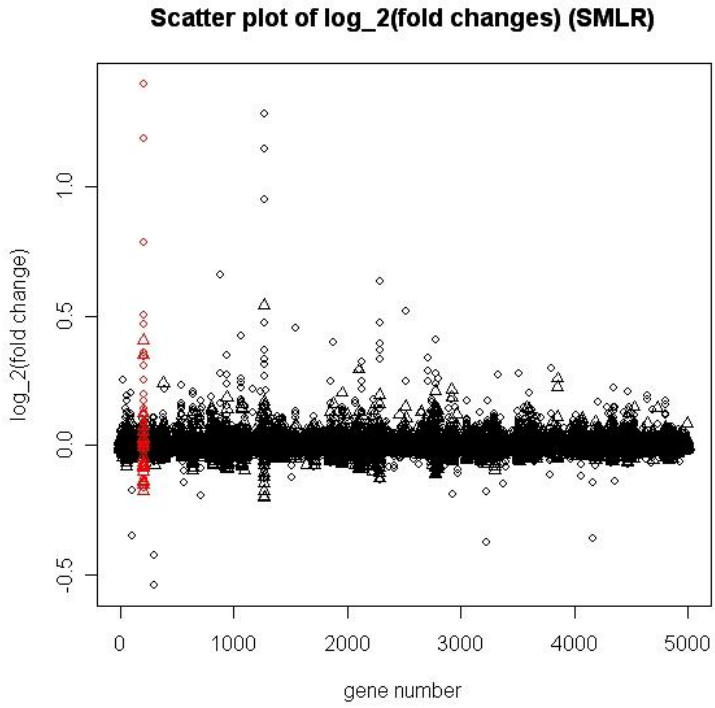

Figure 4.10: Scatter plot of LN-/LN+ log_2(Fold Change data) with genes selected by SMLR.

It is not a surprise to see that this gene is being selected again as it has the highest median fold change. Since only one gene is chosen by this algorithm in this experiment, we cannot comment on its sensitivity to up-regulated or down-regulated genes. However, it is reasonable to suspect that SMLR only selects a small number of genes when the data does not change much over conditions.

Algorithm	gene #			
LASSO	208	546	2751	2802
SLR	208			
SMLR	208			

Table 4.9: Table with the number of all the genes selected by each algorithm using the LN-/LN+ data.

Table 4.9 shows the number of all the genes selected by each algorithm. Only one gene was selected by all of the three algorithms.

4.3 ER+LN+/ER+LN-/ER-LN+/ER-LN- Experiment

After comparing the ability of processing different pairs of the Breast Cancer data data using the three algorithms, the complete Breast Cancer data with all of the four categories (ER+LN+, ER+LN, ER-LN+, ER-LN-) were processed. The fold changes were computed by dividing the replicates from all groups by the average of the replicates from the ER-/LN- group. Because SMLR is the only algorithm that can accommodate data from more than two categories, this experiment focused on its ability in handling this type of data. Also, since the SMLR algorithm we programmed does not have the similar vector length limitation as the LASSO algorithm provided by the R package, the complete data instead of a random sample of 5000 genes was used in this experiement. However, because this data is no longer exactly the same as the data used in the previous sections in this chapter, it is meaningless to reproduce a table similar to Table 3.16 in Section 3.4. Another two tables (Table 4.10 and Table 4.11) with the fold changes of all the genes selected by SMLR in this experiment are provided below.

Similar to the two conditions experiments, there is only a range of tuning parameter values that allows the algorithm to stay converged. Using the tuning parameter 1.5, two genes were selected. It is shown in the median columns of the tables that the two genes are both down-regulated. They both have median fold changes that are small when compared to the ones selected in the previous sections. This may suggest that SMLR has the ability to handle data with more than two categories. Also, it is reasonable to suspect that, because this data does not change much over conditions, only two genes were selected. A further investigation of SMLR in handing data with more than two categories will be provided in the next chapter.

ER-/LN-

Gene #	Gene name	ER-/LN- #1	ER-/LN- #2	ER-/LN- #3
131	AFFX-CreX-3_at	0.97	1.17	1.14
1430	hum_alu_at	1.00	1.11	1.12
		ER-/LN- #4	ER-/LN- #5	ER-/LN- #6
131	AFFX-CreX-3_at	1.00	0.99	1.25
1430	hum_alu_at	0.93	0.93	1.24
		ER-/LN- #7	ER-/LN- #8	ER-/LN- #9
131	AFFX-CreX-3_at	0.88	0.91	0.90
1430	hum_alu_at	0.97	0.86	0.95
		ER-/LN- #10	ER-/LN- #11	ER-/LN- #12
131	AFFX-CreX-3_at	0.83	0.93	1.08
1430	hum_alu_at	0.91	0.99	1.04
		ER-/LN- #13	Median	
131	AFFX-CreX-3_at	0.96	0.97	
1430	hum_alu_at	0.95	0.97	

ER-/LN+

Gene #	Gene name	ER-/LN+ #1	ER-/LN+ #2	ER-/LN+ #3
131	AFFX-CreX-3_at	0.95	1.28	0.93
1430	hum_alu_at	0.83	1.16	0.85
		ER-/LN+ #4	ER-/LN+ #5	ER-/LN+ #6
131	AFFX-CreX-3_at	0.93	0.87	0.83
1430	hum_alu_at	0.87	0.83	0.85
		ER-/LN+ #7	ER-/LN+ #8	ER-/LN+ #9
131	AFFX-CreX-3_at	0.80	0.82	0.86
1430	hum_alu_at	0.83	0.78	0.83
		ER-/LN+ #10	ER-/LN+ #11	Median
131	AFFX-CreX-3_at	0.86	0.86	0.86
1430	hum_alu_at	0.82	1.11	0.83

Table 4.10: Table with the fold changes of a subset of the ER+LN+/ER+LN-/ER-LN+/ER-LN- data that corresponds to the genes selected by SMLR (ER-LN- and ER-/LN+ only).

ER+/LN+

Gene #	Gene name	ER+/LN+ #1	ER+/LN+ #2	ER+/LN+ #3
131	AFFX-CreX-3_at	0.99	0.88	1.00
1430	hum_alu_at	0.89	0.82	0.91
		ER+/LN+ #4	ER+/LN+ #5	ER+/LN+ #6
131	AFFX-CreX-3_at	0.85	0.97	0.92
1430	hum_alu_at	0.79	0.92	0.89
		ER+/LN+ #7	ER+/LN+ #8	ER+/LN+ #9
131	AFFX-CreX-3_at	0.88	0.92	0.91
1430	hum_alu_at	0.95	0.89	0.87
		ER+/LN+ #10	ER+/LN+ #11	ER+/LN+ #12
131	AFFX-CreX-3_at	0.78	0.78	0.81
1430	hum_alu_at	0.88	0.83	0.79
		ER+/LN+ #13	Median	
131	AFFX-CreX-3_at	0.84	0.88	
1430	hum_alu_at	0.79	0.88	

ER+/LN-

Gene #	Gene name	ER+/LN- #1	ER+/LN- #2	ER+/LN- #3
131	AFFX-CreX-3_at	1.87	2.29	0.91
1430	hum_alu_at	1.51	1.99	1.08
		ER+/LN- #4	ER+/LN- #5	ER+/LN- #6
131	AFFX-CreX-3_at	1.11	2.24	0.88
1430	hum_alu_at	1.21	2.30	0.87
		ER+/LN- #7	ER+/LN- #8	ER+/LN- #9
131	AFFX-CreX-3_at	0.89	0.89	0.90
1430	hum_alu_at	0.83	0.83	0.95
		ER+/LN- #10	ER+/LN- #11	ER+/LN- #12
131	AFFX-CreX-3_at	0.83	0.84	0.90
1430	hum_alu_at	0.80	0.78	0.88
		Median		
131	AFFX-CreX-3_at	0.90		
1430	hum_alu_at	0.92		

Table 4.11: Table with the fold changes of a subset of the ER+LN+/ER+LN-/ER-LN+/ER-LN- data that corresponds to the genes selected by SMLR (ER+/LN+ and ER+LN- only).

5 Simulated Data Experiment

After performing some investigations on two real datasets, the algorithms were further tested using simulated fold change data. Through experiments, the performance of each algorithm in selecting regulated genes under samples with various proportions of regulated genes was compared. Also, the ability of each algorithm to identify future treatment samples from a model built using existing data was examined. Moreover, SMLR's ability in handling data with more than two categories was investigated.

The simulated fold change data used in this section was generated from the model created by Mu (2008) in her Master thesis. It was described in her Master thesis that four microarray experimental datasets were used to determine the simulated model and parameters of the fold change data. The four datasets are from an lipopolysaccharide (LPS) time-course microarray experiment. Mu (2008) suggested that the normal distribution is not a good model for the fold change data. However, it was indicated that the base-2 logarithm of the non-regulated fold change data (between 0.5 and 2) approximately follows a normal distribution with the mean of 0 and standard deviation of 0.28, while the uniform distribution may be better than the normal distribution in describing the fold changes of regulated genes. In the LPS time-course microarray experiments data, there were only about 4% of regulated genes at each time point and about 10%-15% in an entire data set. Using the model suggested by Mu (2008) in her Master thesis, the base-2 logarithm transformed fold changes of non-regulated genes were generated from a normal distribution with mean of 0 and standard deviation of 0.28. The fold changes of down-regulated genes were simulated from a uniform distribution in the range of (0.25, 0.5), and the fold changes of up-regulated genes were generated from a uniform distribution in the range of (2, 6). In the following experiments, one sample of each condition was simulated using the model described above. Replicates for a condition were simulated using the normal distribution with the values for that one sample as mean, and a function of the mean as variance. This function of variance was derived from the suggestion proposed by Dudoit et al. (2002) that the logarithm transformed intensity of a gene has approximately constant variance. While fold change is the ratio of the intensities of a gene

under two different conditions, it can be assumed that the logarithm transformed fold change $log\left(\frac{y}{x}\right) = log(y) - log(x)$ also has approximately constant variance. Here, x and y are defined as the intensities of a gene under two different conditions. With the variance of a logarithm transformed fold change being constant,

$$var\left[log(\frac{y}{x})\right] = c, \text{ where } c = \text{constant} \tag{5.0.1}$$

the variance of a fold change can be derived as

$$var\left(\frac{y}{x}\right) = \mu_f^2 c, \text{ where } \mu_f = \text{mean of fold change} \tag{5.0.2}$$

using the Delta method. The Delta method is defined as follows. Let x be a normally distributed random variable with mean μ and variance δ^2, and g be any function. The function $g(x)$ will approximately follow the distribution below,

$$g(x) \dot{\sim} N\left(g(\mu),\ g'(\mu)^2\delta^2\right) \tag{5.0.3}$$

5.1 Experiment #1

In the first experiment, the performance of each algorithm in identifying regulated genes under samples with different proportions of regulated genes was compared. Samples with two different proportions (5% and 15%) of regulated genes were simulated in this experiment. In each of these samples, two conditions of 1000 genes were simulated. While one of the two conditions was control, the treatment condition contained 5% (or 15%) regulated genes. Among all of the regulated genes, half of them were down-regulated, and the other half were up-regulated genes. Twenty replicates were then generated under each condition.

While beginning with the simulated data containing 5% regulated genes, Figure 5.1 is given below to allow an overview of the data.

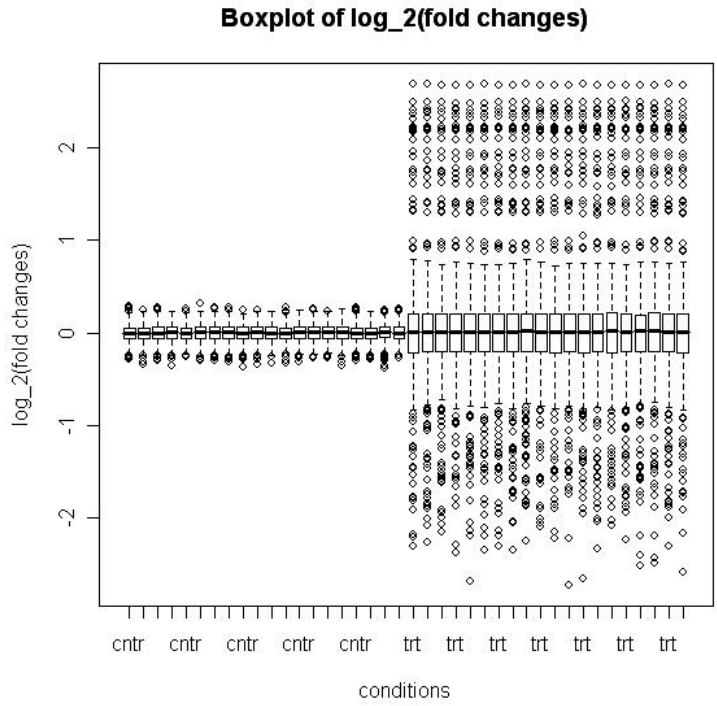

Figure 5.1: Boxplot of log_2(simulated Fold Change data) with 5% regulated genes by replicate.

Since this is simulated data, the boxplot (Figure 5.1) looked slightly different from the ones generated using real data. In this boxplot, the first 20 control condition replicates can be easily distinguished from the last 20 treatment condition replicates. It is known that 5% of the genes in the treatment samples were regulated, and half of them were down-regulated. Figure 5.2 is shown below to provide a better look at the data.

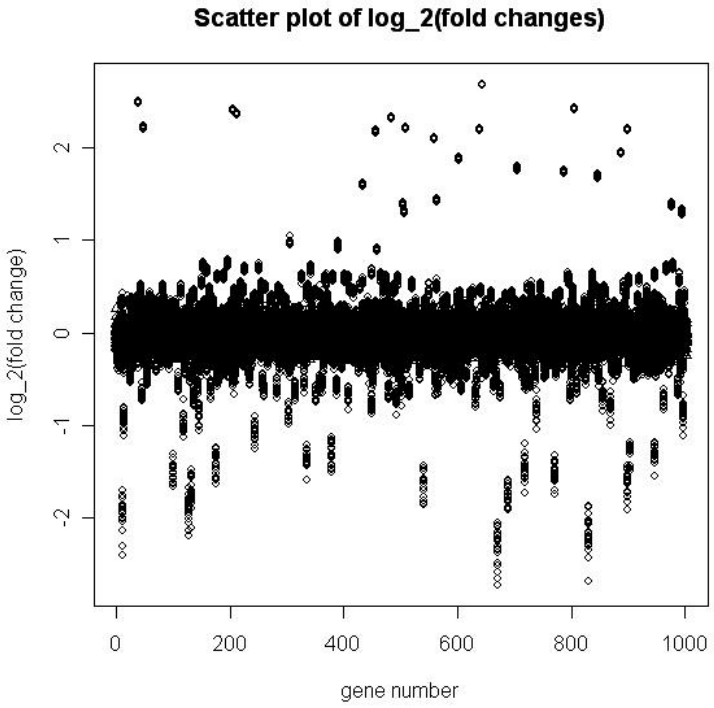

Figure 5.2: Scatter plot of log_2(simulated Fold Change data) with 5% regulated genes by gene number.

Figure 5.2 shows the fold changes of each gene from the 40 replicates; 20 from the control condition, and 20 from the treatment condition. According to the model used to simulate this data, the points that represent the control conditions were most likely to have fold change falling between 0.5 and 2, while in the treatment condition, up-regulated and down-regulated genes may be found. It is known that 5% of all genes (50 genes) are regulated genes, and half (25 genes) are down-regulated. Similar to what we had done in the previous sections using real data, this simulated data was then employed to examine the three algorithms. The data was processed by each of the three algorithms to determine whether they select the genes that are responsible for the most variation.

The first algorithm tested is LASSO; again, there is only a range of tuning parameter values that will allow the algorithm to stay converged. Different tuning parameters were tried, and the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 0.7 was used. With this tuning parameter, 39 genes were chosen. Figure

5.3 is given showing the genes selected in red.

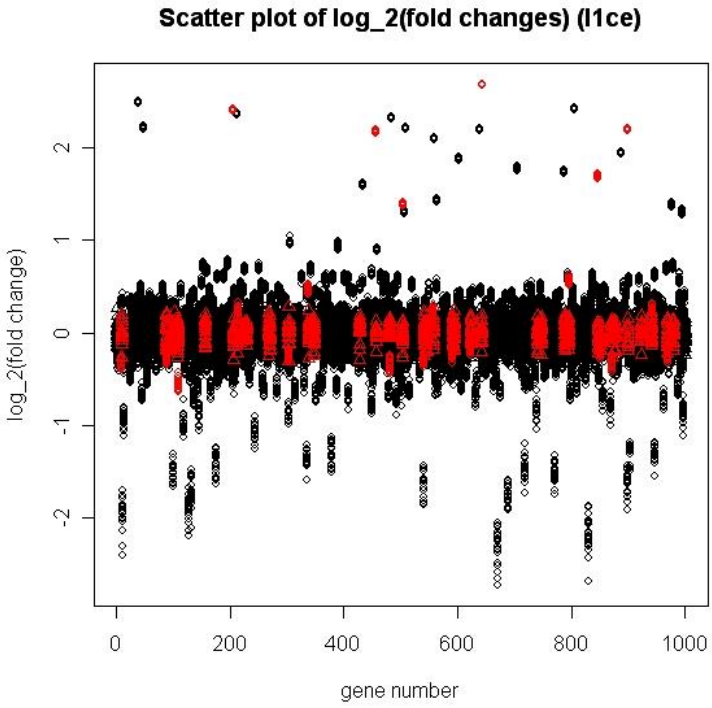

Figure 5.3: Scatter plot of 5% regulated genes log_2(Fold Change data) with genes selected by LASSO.

LASSO	up-regulated genes	down-regulated genes	non-regulated genes
chosen	6	0	33
not chosen	19	25	917

Table 5.1: Table with the number of differentially expressed genes correctly and incorrectly chosen by LASSO using the simulated data with 5% regulated genes.

From Figure 5.3 and Table 5.1 above, we can see only six genes with high fold changes were chosen, while the others had fold changes very close to 1. After looking at the data in detail, it was found that among all the 39 genes selected by LASSO, only six of them are up-regulated (fold change>2), while the rest are not differentially expressed. This result supports the argument given in the previous chapters regarding LASSO being not very accurate on choosing differentially expressed genes. It selected the gene with larger fold changes, but

incorrectly chose some genes that did not change much over conditions, while neglecting some other genes that might possibly be differentially expressed genes. Also, it can be easily seen in the plot that no down-regulated gene was chosen, this again supports the argument given the previous chapters that LASSO is more sensitive to up-regulated genes. Also, after comparing the coefficients calculated using LASSO and the fold changes of the genes chosen, there seems to be no obvious relation between the magnitude of the fold changes and their corresponding coefficients.

The next algorithm we tested using this data was SLR. Again, there is only a range of tuning parameter values that will allow the algorithm to stay converged. Different tuning parameter values were tried, and the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 0.7 was used. Only one gene was selected by SLR. Figure 5.4 that shows the selected gene in red is provided below.

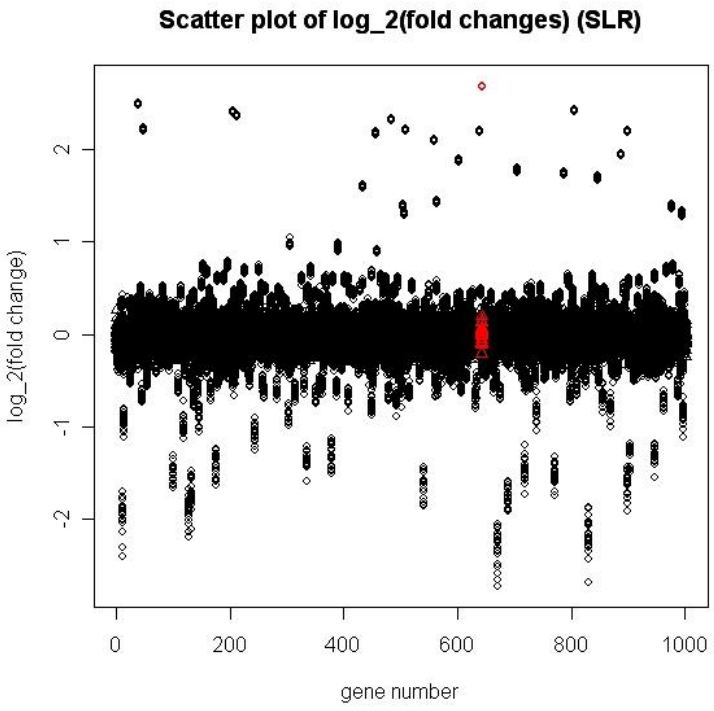

Figure 5.4: Scatter plot of 5% regulated genes log_2(Fold Change data) with genes selected by SLR.

SLR	up-regulated genes	down-regulated genes	non-regulated genes
chosen	1	0	0
not chosen	24	25	950

Table 5.2: Table with the number of differentially expressed genes correctly and incorrectly chosen by SLR using the simulated data with 5% regulated genes.

According to the experience from the previous chapters, it is not a surprise to see SLR selecting only the gene that acted the most differently between the two conditions. When the data was processed with SLR, various tuning parameters had been tried, but the results obtained were identical with only one gene chosen. This once again provides evidence to suspect that the algorithm is not very sensitive to its tuning parameter. The algorithm only selected the gene that contributed to the greatest variation, while neglecting the others that may be meaningful but with comparatively less variation. Again, it is impossible to comment on its sensitivity on up-regulated or down-regulated genes when only one gene was selected by the algorithm.

Finally, the data was processed with the algorithm SMLR. After trying a wide range of tuning parameters, the one that allowed the algorithm to select the most genes was employed. A tuning parameter, 0.7 was used. The algorithm selected ten genes including the one selected by SLR. Figure 5.5 shows the genes chosen in red.

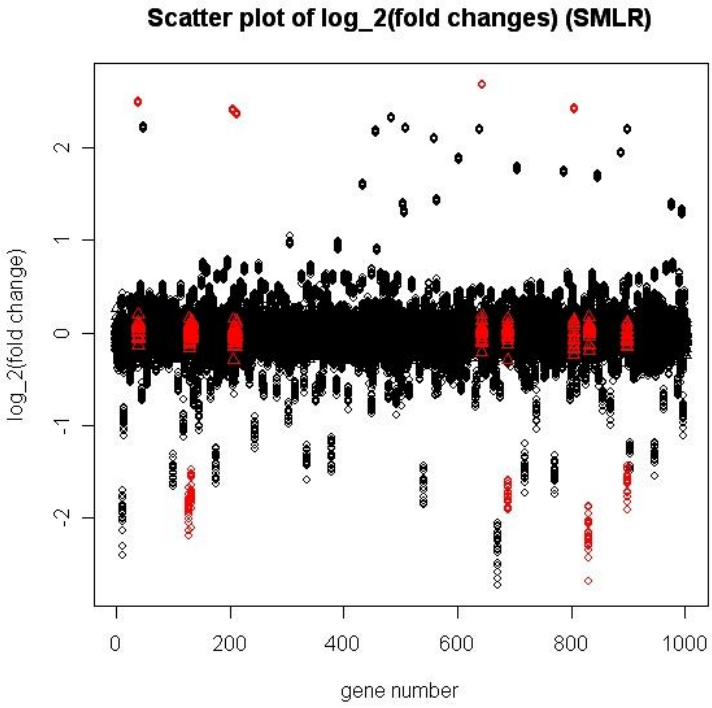

Figure 5.5: Scatter plot of 5% regulated genes log_2(Fold Change data) with genes selected by SMLR.

SMLR	up-regulated genes	down-regulated genes	non-regulated genes
chosen	5	5	0
not chosen	20	20	950

Table 5.3: Table with the number of differentially expressed genes correctly and incorrectly chosen by SMLR using the simulated data with 5% regulated genes.

It is shown in Figure 5.5 and Table 5.3 that all of the genes selected by SMLR are regulated. Among the ten regulated genes, half of them are up-regulated, and the other half are down-regulated. This algorithm is rather accurate in this experiment when comparing to the other two algorithms.

When paying more attention to the coefficients of the multinomial logistic regression model computed using SMLR, it is not difficult to notice a pattern. It seems that SMLR defined coefficients that correspond to up-regulated genes as negative number, while opposite for the

down-regulated genes. The magnitude of the coefficient for each gene was also decided based on the fold change of the gene. The magnitude of the coefficient was larger when the gene varied more between conditions. Figure 5.6 clearly shows this relationship.

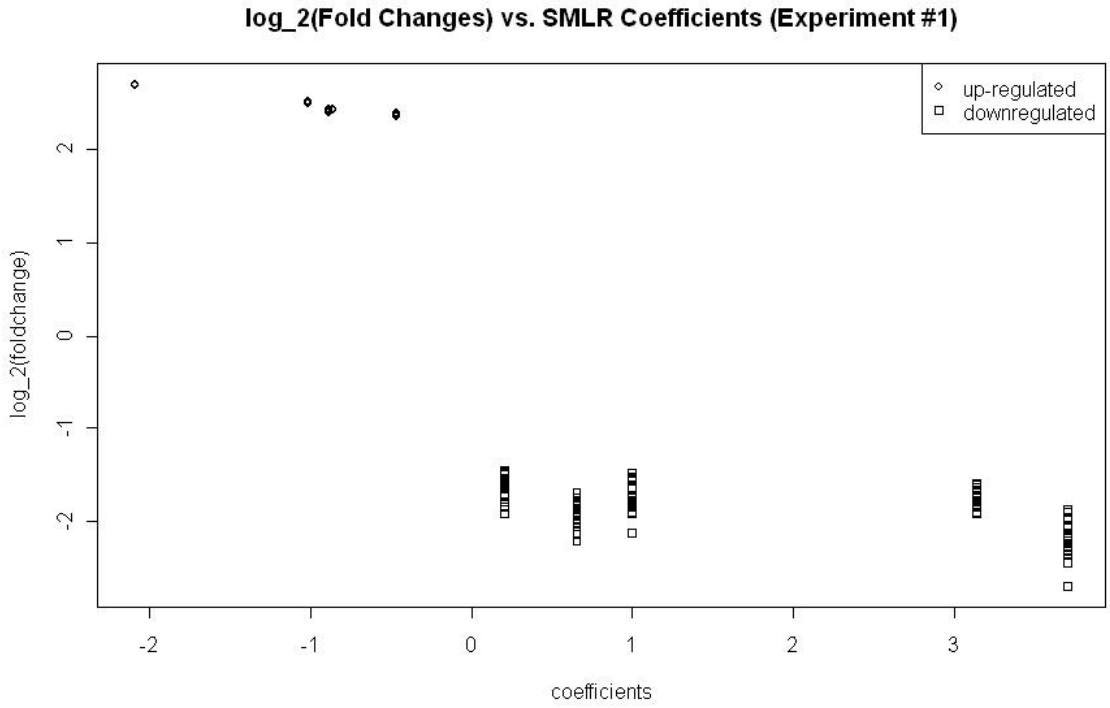

Figure 5.6: Scatter plot of 5% regulated genes log_2(Fold Change data) of selected genes against their corresponding SMLR model coefficients.

In Figure 5.6, fold changes of the genes selected by SMLR were plotted against their corresponding coefficients. It can be seen on the left of the plot that for genes with base logarithm transformed fold changes greater than 1 (up-regulated), their corresponding coefficients are negative. Also, when the fold change of an up-regulated gene is smaller, its corresponding coefficient is closer to 0. Also, it is shown that that the coefficients of the down-regulated genes are positive. Moreover, whenever the fold change of a down-regulated gene is smaller, its corresponding coefficient is larger. This agrees with what was found in the previous chapters using real data.

Since it is known in this simulated experiment that which gene is differentially expressed, Table 5.4 is provided below to summarize the performance of each algorithm when dealing with

this data. Also, true positive rate and false positive rate are also provided in the table. True positive is computed by dividing the number of genes correctly select by the total number of differentially genes, while the false positive rate is calculated by dividing the number of genes incorrectly selected by the total number of genes that are not differentially expressed. It is shown in the table that LASSO did not perform very well in this experiment. It only correctly selected six differentially expressed genes, while selected 33 non-regulated genes. This leads to low true positive rate, and high false positive rate. SLR did not perform very well either, but it did not incorrectly identify non-regulated genes as differentially expressed. Since only one genes is selected by SLR, and the true positive rate is low, while the false positive rate is zero. SMLR performed better when compare to the other two algorithms. It selected ten genes, and all of them were differentially expressed genes. However, because only ten genes were selected, while there are 50 differentially expressed genes in total, the true positive rate is not very high. Since it did not incorrectly identify genes, the false positive rate is zero.

Algorithm	# of genes selected	# Correctly Identified	# Incorrectly Identified	True Positive Rate	False Positive Rate
LASSO	39	6	33	12.0%	3.5%
SLR	1	1	0	2.0%	0.0%
SMLR	10	10	0	20.0%	0.0%

Table 5.4: Summary of the experiment that used the simulated data with 5% regulated genes.

After testing the algorithms using the simulated data with 5% of regulated genes, another simulated data with 15% of regulated genes was employed. Figure 5.7 allows an overview of the data.

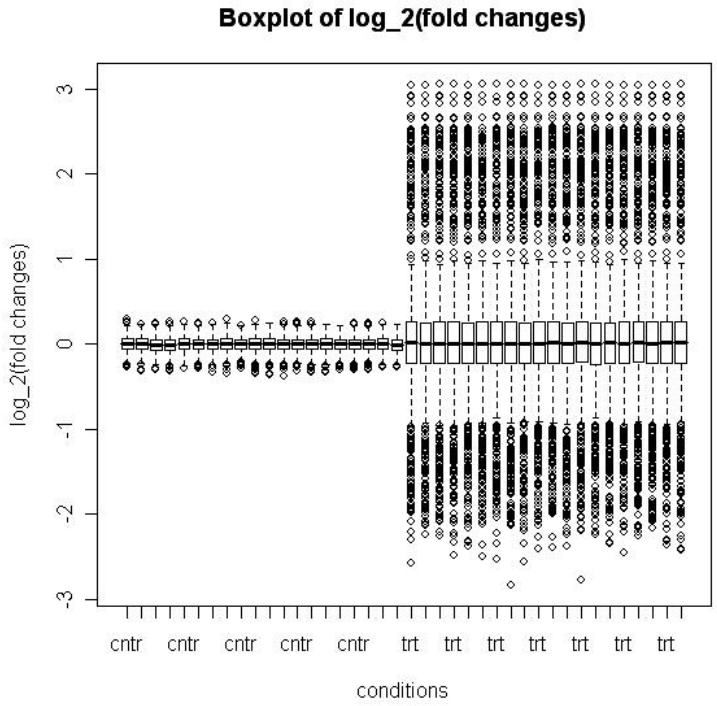

Figure 5.7: Boxplot of log_2(simulated Fold Change data) with 15% regulated genes by replicate.

Again, Figure 5.7 looks slightly different from the one generated using real data as this was built from simulated data. In the figure, the first 20 control condition replicates can easily be distinguished from the last 20 treatment condition replicates. It is known that 15% of the genes in the treatment samples are regulated, with half of them being down-regulated, and the other half being up-regulated. Figure 5.8 provides another look of the data.

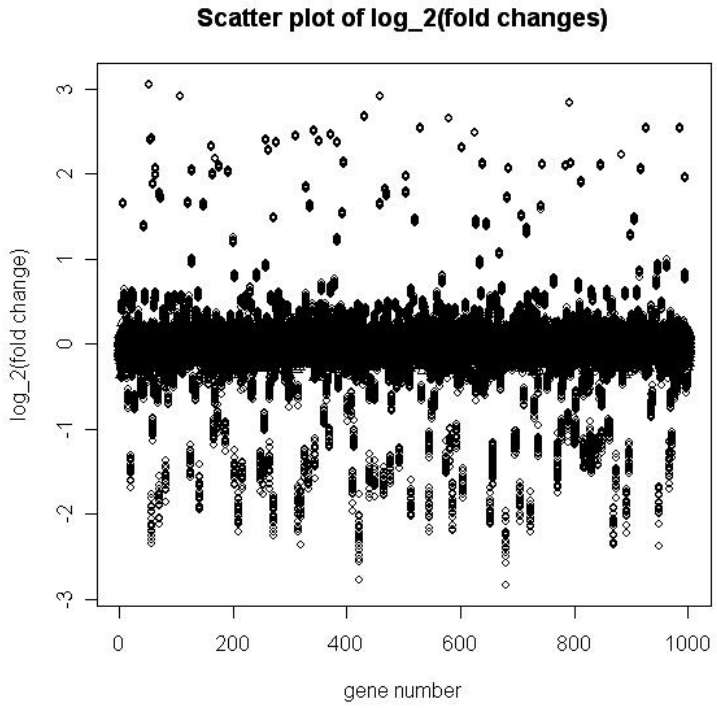

Figure 5.8: Scatter plot of log_2(simulated Fold Change data) with 15% regulated genes by gene number.

While including both conditions, Figure 5.8 shows the fold changes from each gene of all the replicates. It is known that 15% of all genes (150 genes) are regulated genes, and half (75 genes) are down-regulated, and the other half (75 genes) are up-regulated. This simulated data was then employed to examine the three algorithms. The data was processed by each of the three algorithms to allow further investigation on the performance of each algorithm in selecting regulated genes.

The data was first processed using the algorithm LASSO. A tuning parameter, 0.3 was used. With this tuning parameter, 40 genes were chosen. Figure 5.9 is given below showing selected genes in red colour.

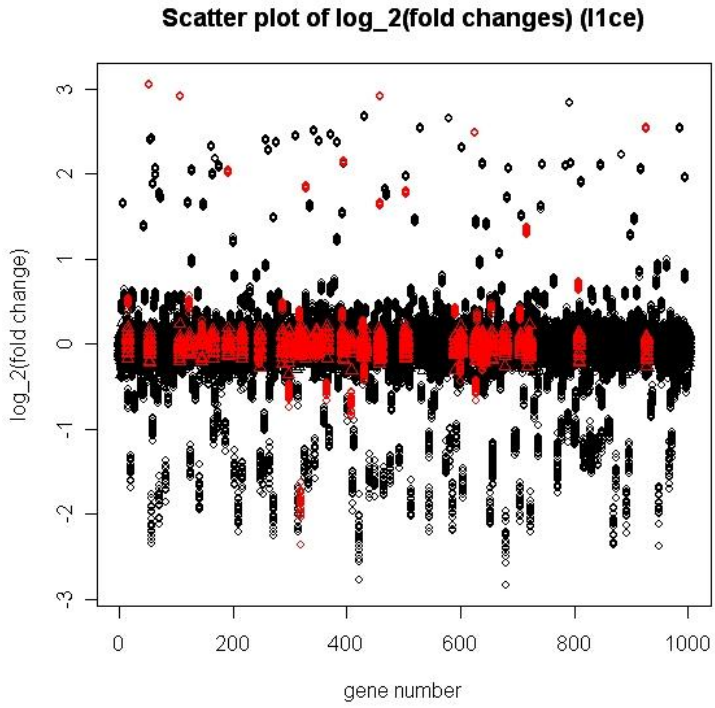

Figure 5.9: Scatter plot of 15% regulated genes log_2(Fold Change data) with genes selected by LASSO.

LASSO	up-regulated genes	down-regulated genes	non-regulated genes
chosen	11	1	28
not chosen	64	74	822

Table 5.5: Table with the number of differentially expressed genes correctly and incorrectly chosen by LASSO using the simulated data with 5% regulated genes.

It is shown in Figure 5.9 and Table 5.5 that, out of the 40 genes selected by LASSO, eleven genes are up-regulated (log_2(fold change)>1), and only one gene is down-regulated (log_2(fold change)<-1). Other than these, the rest selected are not differentially expressed (-1<log_2(fold change)<1). This result once again shows that LASSO did not perform very well in choosing differentially expressed genes. While some genes with large fold changes were selected, some non-regulated genes were selected at the same time. Also, it can be easily seen in the plot that only one down-regulated gene was chosen, this may again suggests that LASSO is more

sensitive to up-regulated genes. Also, after comparing the coefficients calculated using LASSO and the fold changes of the genes chosen, there seems to be no obvious relationship between the magnitude of the fold changes and their corresponding coefficients.

SLR was the next algorithm tested. With the tuning parameter 0.3, only one gene was chosen by SLR. Figure 5.10 shows the selected gene in red colour.

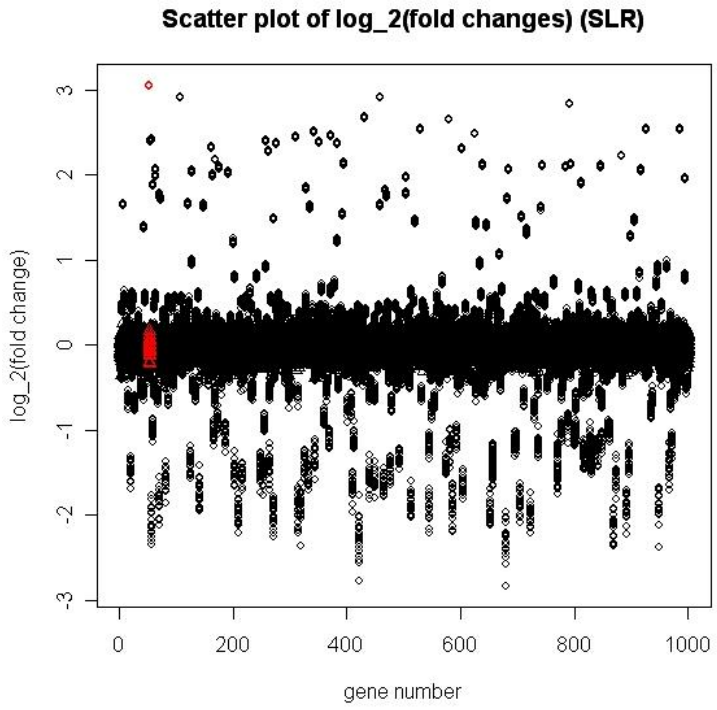

Figure 5.10: Scatter plot of 15% regulated genes log_2(Fold Change data) with genes selected by SLR.

SLR	up-regulated genes	down-regulated genes	non-regulated genes
chosen	1	0	0
not chosen	74	75	850

Table 5.6: Table with the number of differentially expressed genes correctly and incorrectly chosen by SLR using the simulated data with 5% regulated genes.

Similar to the test using the data with 5% of regulated genes, SLR was again selecting only the gene that act the most differently between the two conditions. Regardless of the tuning

parameter used, SLR gave the same result of selecting only the gene that contributed to the greatest variation. This may once again suggest that the algorithm is not very sensitive to its tuning parameter. With only one gene selected, it is impossible to comment on its sensitivity to up-regulated or down-regulated genes.

The last algorithm tested using this data was SMLR. Similar to LASSO and SLR, there is only a range of tuning parameter values that will allow the algorithm to stay converged. A tuning parameter, 0.3 was used. The algorithm selected the 87 genes. With a total of 150 regulated genes in this data, the algorithm was selecting more than half of them. This is a much higher proportion than when this algorithm was processing the data with 5% regulated genes. Figure 5.11 below shows the genes chosen in red colour.

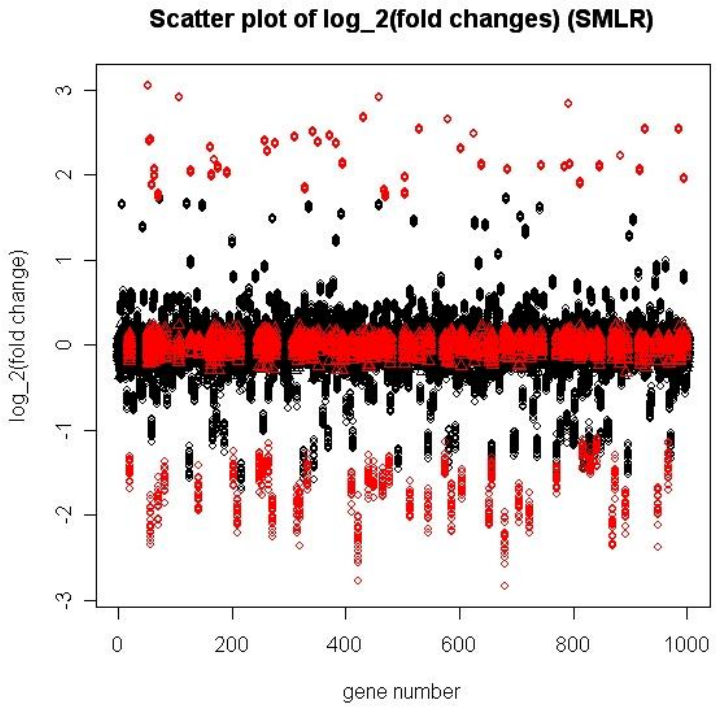

Figure 5.11: Scatter plot of 15% regulated genes log_2(Fold Change data) with genes selected by SMLR.

SMLR	up-regulated genes	down-regulated genes	non-regulated genes
chosen	48	39	0
not chosen	27	36	850

Table 5.7: Table with the number of differentially expressed genes correctly and incorrectly chosen by SMLR using the simulated data with 5% regulated genes.

From Figure 5.11, it is obvious that SMLR selected the group of genes that has the highest, and the lowest fold changes. While details cannot be found in this figure due to high amount of genes selected, a detailed look into the data revealed the fact that all genes selected by SMLR were regulated. It is shown in Table 5.7 that, among the 87 genes chosen, 48 of them were up-regulated genes, and 39 were down-regulated genes.

Once again, a relationship was found between the coefficients of the multinomial logistic regression model computed by SMLR, and the fold changes of the corresponding selected genes. It seems that SMLR defined coefficients that correspond to up-regulated genes as negative number, while opposite for the down-regulated genes. The magnitude of the coefficient for each gene is also decided based on the fold change of the gene. The magnitude of the coefficient is larger when the gene varies more between conditions. Figure 5.12 below clearly shows this relationship.

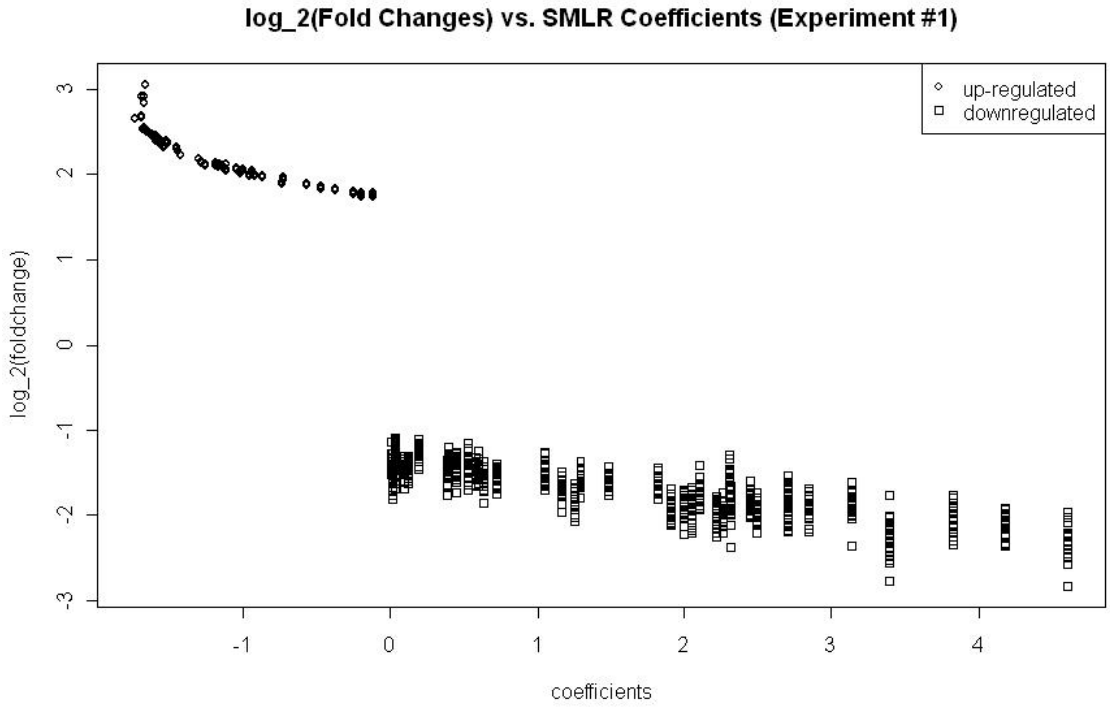

Figure 5.12: Scatter plot of 15% regulated genes log_2(Fold Change data) of selected genes against their corresponding SMLR model coefficients.

In Figure 5.12, base-2 logarithm transformed fold changes of all the genes selected by SMLR was plotted against their corresponding coefficients. It can be seen on the left of the plot that the corresponding coefficients of up-regulated genes are negative. Also, when the fold change of an up-regulated gene is smaller, its corresponding coefficient is closer to 0. The corresponding coefficients for the down-regulated genes selected are positive coefficient. It can also be found in the plot that when the fold change of a down-regulated gene is smaller, its corresponding coefficient is larger. This agrees with what was found in the previous chapters.

Again, since it is known in this simulated experiment that which gene is differentially expressed, Table 5.8 is provided below to summarize the performance of each algorithm when dealing with this data. Also, true positive rate and false positive rate are also provided in the table. It is shown in the table that LASSO performed better when dealing with this simulated data that had a high proportion of differentially expressed genes. It correctly selected 12 differentially expressed genes, while selected 28 non-regulated genes. However,

the true positive rate dropped as this data had a higher proportion of differentially expressed genes. The false positive rate stayed the same. SLR performed similarly when dealing with both data. It only selected one gene that was differentially expressed. Since only one genes is selected by SLR, and the proportion of differentially expressed genes increased, the true positive rate dropped. Again, because it did not identify gene incorrectly, the false positive rate is zero. SMLR improved when dealing with this data. It selected 87 genes, and all of them were differentially expressed genes. With more than half of the differentially expressed genes selected, the true positive rate is quite high at 58%. Again, since it did not incorrectly identify genes, the false positive rate is zero.

Algorithm	# of genes selected	# Correctly Identified	# Incorrectly Identified	True Positive Rate	False Positive Rate
LASSO	40	12	28	8.0%	3.3%
SLR	1	1	0	0.7%	0.0%
SMLR	87	87	0	58.0%	0.0%

Table 5.8: Summary of the experiment that used the simulated data with 15% regulated genes.

With each of the two simulated data (5% and 15% regulated genes) processed by the three algorithms, the performances of each algorithm in selecting genes when employing the two different data were compared. LASSO was the first algorithm investigated. When looking at the genes selected, the proportion of regulated genes does not seem to affect its performance. While selecting a reasonable number of regulated genes, LASSO was more sensitive to up-regulated genes than down-regulated genes using both data. However, it also selected a number of genes that are not differentially expressed. SLR was the algorithm tested after LASSO. Regardless of the regulated genes proportion, only one gene was selected. In both situations, it selected only the gene that varied the most between conditions. Also, in both cases, the tuning parameter was not very sensitive. SLR selected only one gene no matter what tuning parameter we chose. Finally, the performance of SMLR was compared. The algorithm performed very well in both cases by selecting only regulated genes. Also, when similar in both situations, the coefficients computed from the algorithm had close relationship with the fold changes of the corresponding regulated genes. Moreover, when paying attention to the genes selected, we

noticed that SMLR selected a much greater number of genes with the 15% regulated genes data. Although the 15% regulated genes data began with a high proportion of regulated gene (15% vs. 5%), the algorithm selected a much greater number of genes (87 genes vs.10 genes) using this data. This may allow us to suspect that the number of genes selected by SMLR may increase dramatically if the proportion of regulated genes increases in a gene data.

5.2 Experiment #2

After comparing the performance of each algorithm in selecting regulated genes under various proportions of regulated genes, the ability of each algorithm in identifying future treatment samples from a model built using existing data was examined in the second experiment.

Before beginning the second experiment, it may be worthwhile to mention that all three algorithms have the ability to identify which condition a future sample belongs if this future sample is obtained from an experiment that is similar to the one performed when obtaining the data used to build the model. The coefficients calculated in each algorithm can be used to form a regression model that allows direct computation of an indicator, which identifies the condition to which the future sample belongs. The regression models used by the three algorithms vary according to the nature of each algorithm. A brief explanation of how the indicators are calculated using each algorithm is provided below.

Since LASSO was constructed based on an ordinary least squares method, the model used by LASSO is a linear regression model. While letting $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$,

$$y = \sum_j \hat{\beta}_j x_j, \tag{5.2.1}$$

$\hat{\beta}$ is a vector of LASSO coefficients computed. To identify to which condition a sample belongs, the fold changes of the future sample can be plugged into the linear regression model above as x_j's. Together with the LASSO coefficient, a y is calculated for the sample. Since it is defined when building the LASSO model that y=0 when a sample belongs to the control condition, and y=1 when a sample belongs to the treatment conditions, the condition of this new sample can be identified following the same rule. The new sample belongs to the control condition

if the y computed is less than or equal to 0.5, and belongs to the treatment condition if it is greater than 0.5.

SLR used a logistic regression model to identify to which condition a new sample belongs. A logistic function is computed using the coefficients obtained from SLR. While letting $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$,

$$P(y = 1|x) = \frac{e^{\hat{\beta}x}}{1 + e^{\hat{\beta}x}}, \tag{5.2.2}$$

$\hat{\beta}$ is the vector of SLR coefficients computed. This logistic function is also defined as the probability of a sample belonging to a particular condition. To identify to which condition a future sample belongs, the probability of the sample under each condition is computed. The future sample belongs to the condition with the larger computed probability.

SMLR is based on a multinomial logistic model. Similar to SLR, a multinomial logistic function can be used to identify to which condition a new sample belongs. While letting, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$,

$$P(y_i = 1|x) = \frac{e^{\hat{\beta}_i x}}{\sum_j e^{\hat{\beta}_j x}}, \tag{5.2.3}$$

$\hat{\beta}_i$ is the vector of SMLR coefficients compute that corresponds to the condition i. The multinomial logistic function is also defined as the probability that a sample belongs to a particular condition. To identify to which condition a future sample belongs, the probability of the sample under each condition is computed. The new sample belongs to the condition with the largest computed probability.

In this experiment, data from two different conditions (control and treatment) were simulated. Three models were built using each algorithm using different number of replicates (5 replicates, 20 replicates, 40 replicates, and 80 replicates). After that, each model was applied to another 200 samples to compute the percentage that an algorithm correctly identifies the conditions to which these 200 samples belong. Through this, the ability of each algorithm in identifying which condition a future sample belongs using model built was compared. Also, this allows investigation of the question 'Whether a different number of replicates used to build the model affects the ability of the model to identify to which condition a future sample belongs?' To begin, the data with 10% regulated genes data was simulated. The data simu-

lated in this experiment included 1000 genes. Among the 10% of regulated genes, half of them were down-regulated genes, and the other half were up-regulated genes.

Since the purpose of this experiment is not about testing the ability of each algorithm to select genes, the plots similar to the ones given in experiment #1 that show the genes selected are not provided in this experiment. Instead, two tables that summarized the results are provided below.

Algorithm	# of replicates	# of genes selected	# of genes Correctly Identified	# of genes Incorrectly Identified
LASSO	5	10	2	8
	20	20	3	17
	40	40	11	29
	80	79	16	63
SLR	5	1	1	0
	20	1	1	0
	40	1	1	0
	80	1	1	0
SMLR	5	20	12	8
	20	24	7	17
	40	113	68	45
	80	227	96	131

Table 5.9: Summary of the experiment that used different # of training samples.

Table 5.9 summarizes the results of all the three algorithms after building models using different number of replicates (5 replicates, 20 replicates, 40 replicates, and 80 replicates). SLR selected only one gene no matter how many replicates used, and the only gene selected was always differentially expressed. With the increasing number of replicates used, both LASSO and SMLR selected more genes. This caused the number of genes correctly and incorrectly selected to increase as well. The number of genes selected by SMLR increased in a much faster rate than LASSO. This also applied to the number of genes correctly and incorrectly selected. Table 5.10 further illustrates this through true and false positive rates.

Algorithm	# of replicates	True Positive Rate	False Positive Rate
LASSO	5	2.0%	0.9%
	20	3.0%	1.9%
	40	11.0%	3.2%
	80	16.0%	7.0%
SLR	5	1.0%	0.0%
	20	1.0%	0.0%
	40	1.0%	0.0%
	80	1.0%	0.0%
SMLR	5	12.0%	0.9%
	20	7.0%	1.9%
	40	68.0%	5.0%
	80	96.0%	14.6%

Table 5.10: True/False positive rates of the experiments that used different # of training samples.

It is shown in Table 5.10 that the true positive rate for SLR remained to be 1% as the number of correctly selected genes was one no matter how many replicates used. The true positive rate of LASSO increased from 2% to 16% when the number of replicates used increased from 5 to 80. The false positive rate increased at the same time. For SMLR, it is similar to LASSO that the true positive rate increased as the number of replicates used increased. However, the true positive rate of SMLR increased in a much faster rate. It increased from 12% to 96% when the number of replicates used increased from 5 to 80. Also, the false positive rate increased faster than the one obtained from LASSO. After investigating the genes selected, Table 5.11 below shows the corresponding percentage that a model correctly identified to which conditions the 200 samples belong. These models were built using the genes selected.

	# of Training samples			
	5	20	40	80
LASSO	89.5 %	98.5%	99.5%	100.0%
SLR	100.0 %	100.0%	100.0%	100.0%
SMLR	61.5 %	72.5%	88.5%	100.0%

Table 5.11: Percentage of correctly identifying to which conditions the 200 samples belong.

In Table 5.11, we can see that both LASSO and SMLR have increased percentage with the

increased number of replicates used to build the model. SMLR increased more from 61.5% to 100% when compared to LASSO. This may suggest that SMLR relied more on the number of replicates used to build the model to increase its accuracy in identifying in which a future sample belongs, while LASSO being more consistent when different number of replicates used. When looking at SLR, one gene was selected in this experiment no matter how many replicates were used to build the model. This supports the statement made in the previous chapters that SLR is only sensitive to the genes that varied more between conditions, while neglecting others that may possibly be meaningful. With the one gene model, the probability stayed at 100% no matter how many replicates were used.

5.3 Experiment #3

This experiment is very similar to experiment #2. Instead of simulating data with only two conditions, a dataset with four different conditions were simulated. Since only SMLR has the ability to handle data with more than two conditions, SMLR was the only algorithm investigated in this experiment. Its ability to identify future treatment samples from a model built using existing data will be examined. Also, similar to the previous experiment, models were built using different numbers of replicates (5 replicates, 20 replicates, 40 replicates, and 80 replicates). The accuracy of the models built from different number of replicates were then examined.

In this experiment, the data was simulated using the model suggested by Mu (2008) in her Master thesis. It included 1000 genes. In a total of 4 different conditions, one of them is the control, while the others are different treatments. While half of the regulated genes are down-regulated, and the other half are up-regulated, the 15% regulated genes simulated are distributed evenly among the three treatment conditions. Table 5.12 below shows a summary of the results of this experiment. A gene was counted as selected if it was differentially expressed in any of the treatments.

Algorithm	# of replicates	# of genes selected	# Correctly Identified	# Incorrectly Identified
SMLR	5	78	74	4
	20	94	93	1
	40	122	109	13
	80	129	110	19

Table 5.12: Summary of the experiment that used different # of training samples.

It is shown in Table 5.12 that the number of genes selected increase as the number of replicates used increased. This caused the number of genes correctly and incorrectly selected to increase as well. The number of genes selected was 78 with 5 replicates, and increased to 129 with 80 replicated. The number of genes correctly selected increased from 74 to 110. The number of genes selected when using 5 replicates and 20 replicates were high, and did not increase as much as in experiment #2 when the number of replicates increased. Table 5.10 further illustrates this through true and false positive rates.

Algorithm	# of replicates	True Positive Rate	False Positive Rate
SMLR	5	49.3%	0.5%
	20	62.0%	0.1%
	40	72.7%	1.5%
	80	73.3%	2.2%

Table 5.13: True/False positive rates of the experiment that used different # of training samples.

Table 5.13 suggested that the true positive rate increased in a slower rate from 49.3% to 73% when the number of replicates used increased from 5 to 80. Also, the false positive rate increased slowly as well.

	# of Training samples			
	5	20	40	80
Percentage	55.25%	57.8%	80.3%	92.8%

Table 5.14: The percentage of correctly identifying to which condition a future sample belongs.

The performance of SMLR when handling the four conditions data was similar to its

performance when handling the two conditions data in the previous experiments. It is similar to what was found in experiment #2 that the number of genes selected increased with the number of replicates used to build the model. Also, Table 5.14 shows that the percentage of a model correctly identifying a future sample increased dramatically with the increased number of replicates used. The result from this experiment showed that SMLR is capable of handling data with more than two conditions. Also, the more replicates used to build the model, the more accurate it is in identifying to which condition a future sample belongs.

6 Conclusions and Discussion

Microarray analysis is a tool that allows scientists to detect thousands of genes in a small sample simultaneously and to analyze the expression of those genes. From this, a large dataset containing the measurements of the genes' expression levels is produced. Because a microarray is a costly experiment, few replications are usually performed. While considering each gene as one variable, letting p represent the number of variables and n represent the number of replicates, a dataset with $p \gg n$ is generated. Datasets that have $p \gg n$ have always been a challenge for statisticians, while methods such as dimension reduction and variable selection are often used to deal with such problems.

In this study, three different feature selection methods are tested and compared using Microarray data. LASSO is an algorithm that is based on the Ordinary Least Squares method, while SLR is formulated using logistic regression. Due to the nature of these two algorithms, they are built to handle outcomes that take on binary values. SMLR is built using multinomial logistic regression. It has the ability to handle outcomes that take on more than two values.

Two real datasets were used in this study to investigate and compare the ability of the three different algorithms in selecting feature genes. After testing these with both the Intersex data and the Breast Cancer data, we observed that LASSO was capable of selecting differentially expressed genes, however, at the same time, it selected non-differentially expressed genes. We may conclude that LASSO was not very accurate in selecting differentially expressed genes when these microarray datasets were used. It was found that LASSO was more sensitive to up-regulated genes. In contrast to LASSO, SLR was not very sensitive to its tuning parameter when processing the two real datasets. In the experiments, limited amount of differentially expressed genes were selected by SLR regardless of the tuning parameter value used, but it selected the genes that differentiate the most between conditions. SMLR was also tested using the two real datasets. It was quite sensitive to its parameter, and it selected a reasonable number of genes. Among the genes selected, a higher proportion were differentially expressed genes when compared to LASSO. Also, the model coefficients generated by SMLR had a close relationship with the magnitude of fold changes. We may conclude that SMLR performed the

best in selecting feature genes when dealing with these two real datasets. After comparing the performance of the three different algorithms in selecting feature genes using data under two conditions, SMLR's ability to work with data that had more than two conditions was tested. It was found that SMLR was capable of selecting differentially expressed genes even with data that had more than two conditions. It selected the genes that were also identified as differentiated genes when the datasets were separated into different pairs of conditions.

After testing the three algorithms with the two real datasets, they were again tested using simulated data. Since differentially expressed genes were known in the simulated data, the ability and accuracy of algorithms to select differentially expressed genes was verified by computing the true positive and false positive rates. The results from the experiments using simulated data agreed with what was found when testing with the real datasets. In addition, simulated data with different proportions of differentially expressed genes were used for further testing. It was found that the proportion of differentially expressed genes in the data did not affect the performance of LASSO and SLR. This is quite similar to the results of the correspondence analysis method suggested by Mu (2008). However, SMLR behaved differently. The number of genes selected by SMLR increased dramatically when the proportion of regulated genes increased in the simulated data. Moreover, the ability of the algorithms to identify future treatment samples was investigated and compared. During the experiments, SMLR performed the best by providing the highest percentage of correctly identified conditions to which a future sample belongs. As the number of replicates used to build the model increased, the number of genes selected by SMLR increased dramatically when compared to the other algorithms. This also applied to the number of genes correctly and incorrectly selected.

Overall, it was found that SMLR performed the best in both feature selection, and future treatment sample identification when using the microarray data. However, further analyses and investigations may be done to improve and extend this study in the future. First of all, in this study, due to limitations on the amount of data, tuning parameter values were defined as the values that allowed the algorithms to select the most genes. There are other methods such as cross-validation, which may enable the algorithms to select tuning parameter

values that generate more efficient models based on the prediction accuracy. Also, the genes of the breast cancer data used in this study did not change much over conditions. A few more datasets should be employed from different sources to further examine the algorithms. Furthermore, since we were comparing three algorithms in this study, and both LASSO and SLR were not functioning well when handing base-2 logarithm transformed data, fold change data without transformation were used in all the experiments. It may be interesting to examine the performance of SMLR with base-2 logarithm transformed data. Further investigation may also be done on SMLR with microarray data that does not have the size limitation of 5000 genes. In addition, an iterative method can be employed to increase the strength of SMLR in selecting feature genes. When SMLR did not select the target amount of feature genes during the first round of processing, it can be applied again on the original data with the selected genes removed. This can be repeated until the amount of genes selected are satisfactory. Similarly, if SMLR selected too many genes in the first round of processing, it can be applied on the genes selected, and further identify the genes that contribute to more variation. This can be repeated until the number of genes selected are satisfactory.

7 References

Böhning, D., 1992: Multinomial Logistic Regression Algorithm. *Annals of the Institute of Statistical Mathematics*, **44**, 197-200.

Dudoit, S., Yang, Y., Callow, M., and Speed, T., 2002: Statistical Methods For Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments: *Statistica Sinica*, **12**, 111-139.

Hageman, L.A. and Young, D.M., 1981: *Applied Iterative Methods*. New York: Academic Press.

Krishnapuram, B., Carin, L., Figueiredo, M.A., and Hartemink, A., 2005: Sparse Multinomial Logistic Regression: Fast Algorithm and Generalization Bound: *IEEE Transactions on Pattern Analysis and Machines Intelligence*, **24**, 957-968.

Mu, R., Applications of Correspondence Analysis in Microarray Data Analysis (M.A. thesis, University of Victoria, 2008).

R Development Core Team, 2008: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Shevade, S.K. and Keerthi, S.S., 2003: A Simple and Efficient Algorithm for Gene Selection Using Sparse Logistic Regression. *Bioinformatics*, **19**, 2246-2253.

Tibshirani, R., 1996: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, Jr., J.A., Marks, J.R., and Nevins, J.R., 2001: Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles. *Proceedings of the National Academy of Sciences*, 11462-11467.