Doubly Sparse Regularized Regression Incorporating Graphical Structure Among Predictors

by

Matthew Stephenson

A Thesis presented to The University of Guelph

In partial fulfilment of requirements for the degree of Doctor of Philosophy in Mathematics and Statistics

Guelph, Ontario, Canada

© Matthew Stephenson, December, 2018
Recent research has demonstrated that information learned from building a graphical model on the predictor set of a regularized linear regression model can be leveraged to improve prediction of a continuous outcome. This thesis proposes the doubly sparse regression incorporating graphical structure among predictors (DSRIG) model, and its logistic regression counterpart, doubly sparse logistic regression incorporating graphical structure among predictors (DSLRIG). In general, the regularization scheme of these models works by building an undirected graph over the predictor set and then using the resulting neighbourhoods of the graph to form a set of (overlapping) groups. Sparsity is encouraged both within and among the groups that contribute to the overall estimation of the regression parameters. Together, DSRIG and DSLRIG provide a unified framework for the fitting of many other commonly used regularization schemes.

In this thesis, a combination of simulation and analysis of real world data are used to evaluate and compare model performance. Ultimately, the DSRIG and
DSLRIG models improve outcome prediction and parameter estimation compared to previously proposed methods. A finite sample error bound is derived for DSRIG in the case of a quantitative outcome and predictors distributed as multivariate normal. Guidelines for the implementation of DSRIG in the analysis of real world data are also provided.
Acknowledgements

First and foremost, I would like to thank my advisors Dr. Ayesha Ali and Dr. Gerarda Darlington. Without them I would never have found myself in the field of statistics. Their support, encouragement, and guidance throughout my graduate studies have made this thesis possible.

I would like to express my gratitude to my advisory committee member Dr. Flavio Schenkel for providing useful conversations and the data used as a motivating example in this thesis. I would like to extend my sincere appreciation to Dr. James Squires for helping to bridge the gap between prior scientific findings and the computational analyses presented in this thesis.

I would also like to thank Dr. Xin Gao, Dr. Zeny Feng, and Dr. Julie Horrocks, for taking the time to be part of my examining committee. To all the other members of the Department of Mathematics and Statistics at the University of Guelph, thank you for making this process so enjoyable.

I am grateful to both the National Sciences and Engineering Research Council and the Ontario Graduate Scholarship for providing funding to allow me to pursue my graduate studies.

Finally, thank you to all my friends and family for their support throughout my graduate studies.
Table of Contents

Abstract ii
Acknowledgements iv
List of Tables vii
List of Figures ix
1 Introduction 1
2 Background 4
  2.1 Linear regression and regularized linear regression 4
  2.2 Logistic regression and regularized logistic regression 6
  2.3 Link between linear and logistic regression parameters 7
  2.4 Graphical models 9
    2.4.1 Gaussian graphical models 11
  2.5 Regularized regression incorporating graphical structure 12
    2.5.1 Sparse regression incorporating graphical structure among predic-
    tors: SRIG 13
    2.5.2 Limitations of SRIG and thesis overview 16
3 Doubly Sparse Regression Incorporating Graphical Structure Among
  Predictors 19
  3.1 Abstract 19
  3.2 Introduction 20
  3.3 Methods and Motivation 22
    3.3.1 SRIG 23
    3.3.2 DSRIG 25
    3.3.3 Estimation by Proximal Gradient Descent 26
  3.4 Estimator Properties 28
    3.4.1 Properties of $\mathcal{R}(\beta)$ and $\mathcal{L}(\beta)$ 31
    3.4.2 Finite Sample Error Upper Bound 33
  3.5 Simulation Study 34
    3.5.1 Simulation Study Design 35
    3.5.2 Simulation Study Results 38
  3.6 Application 40
    3.6.1 Alzheimer’s Disease Data 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6.2</td>
<td>Blood Brain Barrier Data</td>
<td>44</td>
</tr>
<tr>
<td>3.6.3</td>
<td>Results of Data Analyses</td>
<td>45</td>
</tr>
<tr>
<td>3.7</td>
<td>Discussion and Future Work</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td>DSRIG: Incorporating Graphical Structure in the Regularized Modeling of SNP Data</td>
<td>51</td>
</tr>
<tr>
<td>4.1</td>
<td>Abstract</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Background</td>
<td>52</td>
</tr>
<tr>
<td>4.3</td>
<td>Methods</td>
<td>55</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Doubly Sparse Regression Incorporating Graphical Structure Among Predictors</td>
<td>55</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Cross-Validation for Genetic Selection</td>
<td>57</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Empirical Evaluation</td>
<td>59</td>
</tr>
<tr>
<td>4.4</td>
<td>Results</td>
<td>62</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>DSLRIG: Leveraging Predictor Structure in Logistic Regression</td>
<td>69</td>
</tr>
<tr>
<td>5.1</td>
<td>Abstract</td>
<td>69</td>
</tr>
<tr>
<td>5.2</td>
<td>Introduction and Background</td>
<td>70</td>
</tr>
<tr>
<td>5.3</td>
<td>Notation and Methods</td>
<td>72</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Notation</td>
<td>72</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Methods</td>
<td>73</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Boar Taint Data</td>
<td>76</td>
</tr>
<tr>
<td>5.4</td>
<td>Simulation Study</td>
<td>77</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Data Generation</td>
<td>77</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Model Fitting</td>
<td>78</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Results</td>
<td>80</td>
</tr>
<tr>
<td>5.5</td>
<td>Application</td>
<td>82</td>
</tr>
<tr>
<td>5.6</td>
<td>Discussion</td>
<td>87</td>
</tr>
<tr>
<td>6</td>
<td>Discussion, Conclusions and Future Work</td>
<td>90</td>
</tr>
<tr>
<td>6.1</td>
<td>Analysis of Boar Taint Data</td>
<td>92</td>
</tr>
<tr>
<td>6.2</td>
<td>Future Work</td>
<td>94</td>
</tr>
</tbody>
</table>

Bibliography 96

A Supplementary Information for Doubly Sparse Regression Incorporating Graphical Structure Among Predictors 101

B Additional Results to Chapter 5 110

C Comparison of Boar Taint Analyses 115
List of Tables

3.1 The \( \ell_2 \)-distance and relative prediction error of model fits under two different cross-validation splits for training/validation/test for random and scale-free graphs. .............................................. 38

3.2 Proportion of runs each method had the lowest \( \ell_2 \)-distance or relative prediction error when the predictor graph is estimated from the training data. The median difference between SRIG and DSRIG for the \( \ell_2 \)-distance and RPE is shown in brackets. ......................... 40

3.3 Average mean square error in prediction and mean and standard deviation for number of non-zero regression coefficients for 90 permutations of the ADNI and blood brain barrier data. ......................... 45

3.4 Predictors found to be non-zero in at least 80 models for the ADNI data. 47

4.1 List of 20 edges from the consensus graph for the Duroc breed. Node labels, as shown in Figure 4.2, and genes associated with the edge ends are also provided. .................................................. 62

4.2 Average mean square error in prediction, mean number of non-zero regression coefficient estimates (standard deviation shown in brackets) and median standard deviation of the regression coefficients (maximum shown in brackets) for three breeds of pig and two compounds across 90 permutations. .................................................. 63

4.3 Summary across the 90 estimates of the unstandardized DSRIG regression coefficients for the prediction of androstenone in the Duroc breed restricted to the seven SNPs that had the largest absolute median values across the 90 permutations. ............................ 66

5.1 Median proportion of observations misclassified in the test set across 100 iterations for ten different sample sizes of training and validation sets. .................................................. 81

5.2 Comparison of classification accuracy (tainted: yes/no), median standard deviation of the cross-validated regression coefficients and median and interquartile range (IQR) for number of non-zero coefficient estimates for three breeds of pigs. .............................................. 85

5.3 Summary across the 513 estimates of the DSLRIG (max \( \| \beta \|_2 \)) regression coefficients for the Duroc breed restricted to the five covariates that had the largest absolute median values. ............................ 87
B.1 Median proportion of observations misclassified in the test set across 50 iterations for five different sample size of training and validation sets for regression parameters on original scale. 113
List of Figures

2.1 Comparison of the cumulative distribution function for the logistic distribution and normal distribution both with mean 0 and variance 0.85035. 9

2.2 Undirected graphical model with 6 nodes and 9 edges where each node is of degree 3. 11

3.1 (a) Undirected graph representing the joint distribution over \{Y, X_1, X_2, X_3\}, and (b) predictor graph over \{X_1, X_2, X_3\} obtained after marginalizing Y out of the joint distribution in (a). 24

3.2 Estimated predictor graphs from independent data sets generated from the same scale-free parent graph. 36

3.3 Difference (SRIG-DSRIG) for 100 simulations from each of 30 random parent graphs under Scenario 2 with sample sizes of 40/40/400 for (a) $\ell_2$-distance and (b) RPE. Points above the zero line indicate when DSRIG performed better. 41

3.4 Estimated predictor graph for a single training segmentation of the ADNI data. 43

3.5 Difference in MSPE (SRIG - DSRIG) for 90 permutations of ADNI data. Points above the zero line indicate when DSRIG performed better. 46

4.1 Normal quantile-quantile plot of the log of skatole for the Duroc breed. 60

4.2 Undirected consensus graph for the (a) Duroc and (b) Yorkshire breeds. Nodes, labelled V1, ... , V103, represent the 103 candidate SNPs. Nodes V1, V2, and V3, corresponding to the HSD3B1 gene and its promoter, are circled. Nodes V19 and V20, corresponding to the CYB5B gene, are boxed. 61

4.3 Difference in MSPE (LASSO-DSRIG) across 90 permutations for all breed and compound combinations. Positive values represent permutations in which the DSRIG outperformed the LASSO with respect to MSPE. 64

4.4 Boxplot of regression coefficient estimates for 65 SNPs across 90 permutations for the DSRIG method. Three SNPs corresponding to the HSD3B1 gene and its promoter, are marked above boxplot by a triangle and two SNPs corresponding to the CYB5B gene are marked above boxplot by a square. 65
5.1 (a) Median proportion of the true non-zero regression coefficients correctly estimated to be non-zero; (b) Overall match rate, given by the median proportion of correctly classified coefficients (zero or non-zero). Proportions based on 100 iterations for ten different sample sizes... 81

5.2 Median mean absolute rank difference of the absolute value of the coefficients over the true non-zero regression coefficients in (a); and all coefficients in (b).  

5.3 Undirected predictor graph for the Duroc breed from one of the data splits.  

B.1 Histogram of estimated probabilities for an outcome of interest for regression coefficients multiplied by 1 in (a) and 3 in (b). 111

B.2 (a) Median proportion of the true non-zero regression coefficients correctly estimated to be non-zero; (b) Overall match rate, given by the median proportion of correctly classified coefficients (zero or non-zero). Proportions based on 50 iterations for five different sample sizes and regression parameters on original scale.  

B.3 Median mean absolute rank difference of the absolute value of the coefficients over the true non-zero regression coefficients in (a); and all coefficients in (b). Medians calculated over 50 iterations for five different sample sizes and regression parameters on original scale.  

C.1 Boxplots of cross-validated regression parameter estimates across 65 SNPs for prediction of androstenone expression in the Duroc breed; upper panel corresponds to DSRIG estimates for androstenone treated as quantitative; lower panel corresponds to DSLRIG-max∥ \hat{\beta} ∥_2 estimates for dichotomized androstenone. The SNPs are ordered the same in both plots and a similar set of SNPs are associated with large median regression coefficient estimates and boxplots bounded away from zero. SNPs in the HSD3B1 gene and its promoter are found in boxplots 1 through 3 (in blue box) and SNPs in the CYB5B gene are found in boxplots 6 and 7 (in red box).  

116
Chapter 1

Introduction

This thesis contributes to the areas of statistical learning and regularized regression. Statistical learning encompasses a broad range of modelling techniques where often the objective is the prediction of an outcome, $Y$, from a set of input features $X$ (Hastie et al., 2001). As data collection has become cheaper, data sets continue to increase in size. However, large data sets can present challenges to conventional parameter estimation particularly when the number of parameters, $p$, exceeds the sample size, $n$, or when multicollinearity exists among the set of predictors. In either case, the predictor matrix $X$ will not be of full rank and the regression parameter estimates may not be unique (Hastie et al., 2001). Furthermore, when the number of predictors is large, identification of predictors that are associated with the outcome can form a secondary research objective.

Regularization presents one method to allow for valid parameter estimation even when $X$ is not of full rank. Recent advances in regularized regression have demonstrated that incorporating the graphical structure of the predictors into the regularization scheme may improve the predictive model (Li and Li, 2008; Yang et al., 2012; Yu and Liu, 2016). For example, the Sparse Regression Incorporating
Graphical structure among predictors (SRIG) (Yu and Liu, 2016) leverages information from building an undirected graph over the predictors. The neighbourhoods of the predictor graph are used to form a set of (overlapping) groups and sparsity is encouraged among the groups that contribute to the overall estimation of $\beta$ through an $\ell_2$-groupwise penalty. However, it may be interesting to allow for greater flexibility and further encourage within group sparsity. To that end, this thesis explores the effects of adding an additional level of sparsity and presents a novel regularization scheme.

The objectives of this thesis are as follows:

1. To propose and evaluate the doubly sparse regression incorporating graphical structure among predictors (DSRIG) model;

2. To derive a finite sample error bound for the DSRIG model for the case of a quantitative outcome and predictors distributed as multivariate normal (MVN);

3. To provide guidelines for the implementation of DSRIG in the analysis of real world data;

4. To propose and evaluate the doubly sparse logistic regression incorporating graphical structure among predictors (DSLRIG) model, which is an extension of DSRIG to the binary outcome setting.

The two models proposed in this thesis, DSRIG and DSLRIG, result in improved predictive and model selection accuracy while retaining flexibility. Together,
they provide a unified framework for the fitting of many other commonly used regularization schemes. A combination of simulation and analysis of real world data are used to evaluate and compare model performance.

The organization of this thesis is as follows. Chapter 2 provides background on linear, logistic and regularized regression, graphical models and the sparse regression incorporating graphical structure among predictors (SRIG) model (Yu and Liu, 2016). Chapter 3 presents the new DSRIG model in the case of a quantitative outcome and MVN predictors. Chapter 4 details an application of DSRIG in the modelling of genotype-phenotype data when the predictor set is comprised of single nucleotide polymorphism (SNP) genotypes encoded as $\{0, 1, 2\}$. DSRIG is shown to be a useful model for these data despite the violation of the MVN assumption. Chapter 5 presents DSLRIG, which extends the DSRIG model to the binary outcome setting. Chapter 6 then discusses overall conclusions and areas for future work.
Chapter 2

Background

2.1 Linear regression and regularized linear regression

Linear regression is commonly used to estimate the relationship between an outcome and its input features. Consider the classic linear model:

\[ Y = X\beta + \epsilon \]  \hspace{2cm} (2.1)

where \( Y \) is an \( n \times 1 \) response vector, \( X \) is an \( n \times p \) matrix of predictors, \( \beta \) is a \( p \times 1 \) vector of regression coefficients and \( \epsilon \) is an \( n \times 1 \) vector of independent and identically normally distributed random variables with mean 0 and variance \( \sigma^2 \). The ordinary least squares (OLS) regression parameter estimates may be found as the solution to:

\[
\arg\min_{\beta} ||Y - X\beta||^2_2,
\]  \hspace{2cm} (2.2)

where \( || \cdot ||_p \) is used to denote the \( \ell_p \)-norm. This optimization problem corresponds to the minimization of the sum of squares residual (SSR); that is, the sum of the squared distances between the observed values of \( Y \), and their predicted values \( \hat{Y} = X\hat{\beta} \).
The solution to Equation (2.2), $\hat{\beta}$, can be found analytically through matrix multiplication:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$  \hspace{1cm} (2.3)

As the estimation of the regression parameters requires the inversion of $(X'X)$ it is necessary that the matrix $X$ be of full rank to ensure that $\hat{\beta}$ is unique (Hastie et al., 2001). However, this condition is violated in the presence of multicollinearity as well as in the case of high dimensional data where $p > n$. In the high dimensional case, two common approaches can be employed. The first approach involves dimension reduction using techniques such as principal component analysis (Hastie et al., 2001). The second approach, and main focus of this thesis, involves regularization of the regression coefficients (Hastie et al., 2001).

Regularized regression has been proposed as one method to improve predictive accuracy of OLS by addressing the bias-variance tradeoff and can provide parameter estimates even when $X$ is not of full rank (Hastie et al., 2001). While OLS involves choosing the regression parameters that minimize the SSR, regularized regression models additionally minimize a penalty term, $R(\beta)$, weighted by a tuning parameter, $\lambda$, $\lambda \geq 0$. The parameter estimates of a regularized regression model can be found as the solution to an optimization problem of the form:

$$\arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda R(\beta).$$  \hspace{1cm} (2.4)

The parameter $\lambda$ controls the amount of regularization and when $\lambda$ takes on a value
of 0, regularized regression is equivalent to OLS. The penalty term $\mathcal{R}(\mathbf{\beta})$ is typically a norm of the regression parameter vector. One common example of a norm-based penalty is the $\ell_1$ penalty, $\sum_{i=1}^{p} |\beta_i|$, of the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) with parameter estimates found as the solution to:

$$\arg \min_{\mathbf{\beta}} ||\mathbf{Y} - \mathbf{X}\mathbf{\beta}||_2^2 + \lambda ||\mathbf{\beta}||_1. \quad (2.5)$$

In addition to shrinking coefficients towards zero, the LASSO may also perform variable selection by setting some regression coefficients precisely to zero.

### 2.2 Logistic regression and regularized logistic regression

If instead $\mathbf{Y}$ represents a binary outcome, then the linear regression model is no longer appropriate and the logistic regression model is preferred. Let $k = 1, \ldots, n$ index the individual observations. Then the logistic regression model may be defined as:

$$g_\beta(\mathbf{X}_k) = \log \left( \frac{P(Y_k = 1|\mathbf{X}_k)}{1 - P(Y_k = 1|\mathbf{X}_k)} \right) = \mathbf{X}_k \mathbf{\beta}. \quad (2.6)$$

Logistic regression involves the use of the logit link function, $g_\beta(\mathbf{X}_k)$, to result in a model where the log odds are a linear function of the predictors. Equation (2.6) can be rearranged to solve for $P(Y_k = 1|\mathbf{X}_k)$ as:

$$\pi_k = P(Y_k = 1|\mathbf{X}_k) = \frac{1}{1 + \exp(-\mathbf{X}_k \mathbf{\beta})}. \quad (2.7)$$
Parameter estimates for the logistic regression model can then be found as the solution to:

$$\arg \min_{\beta} \{-\ell(\beta)\} = \arg \min_{\beta} \left\{ -\sum_{k=1}^{n} [y_k \log(\pi_k) + (1 - y_k) \log(1 - \pi_k)] \right\}, \quad (2.8)$$

where $\ell(\beta)$ is the log likelihood function. Since there is no typically closed form solution for the regression parameter estimates, iterative methods such as the Newton-Raphson method or gradient descent may be used for parameter estimation.

When $X$ is not of full rank, parameter estimates for the logistic regression model will not be valid. Similar to the linear regression case, this may be addressed through the addition of a penalty term with estimates found as the solution to:

$$\arg \min_{\beta} \{-\ell(\beta) + \lambda R(\beta)\}. \quad (2.9)$$

The $\ell_1$ penalty of the LASSO (Tibshirani, 1996) again forms a common choice.

### 2.3 Link between linear and logistic regression parameters

Following Moser and Coombs (2004), let $Y$ be a quantitative response resulting from the classical linear model defined in Equation (2.1) but where the underlying distribution for the $n$ components of $\epsilon$ is the logistic distribution with mean 0 and
variance $\sigma^2 > 0$. Then let $Y^*$ be a dichotomized version of the response such that:

$$Y^*_k = \begin{cases} 
1 & \text{for } Y_k \geq c \\
0 & \text{for } Y_k < c
\end{cases}$$

(2.10)

where $c$ is an arbitrary threshold. Let $\theta$ represent the unknown parameter vector for the logistic regression model of the dichotomized response, i.e.,

$$\log \left( \frac{P(Y^*_k = 1|X_k)}{1 - P(Y^*_k = 1|X_k)} \right) = X_k \theta.$$  

(2.11)

Then it can be shown that:

$$\theta = \frac{\rho \beta}{\sigma},$$

(2.12)

where $\rho = \frac{\pi}{\sqrt{3}}$ (Moser and Coombs, 2004).

Although the above assumes that the distribution of the errors is logistic, there is a close proximity between the logistic and normal distributions. Figure 2.1 plots a comparison of the cumulative distribution functions for the logistic and normal distributions with mean 0 and variance 0.850 (parameter settings used for simulation of the error component in Chapter 5). Given the close proximity of these distributions, it is reasonable to expect that the identities and rank of the non-zero regression coefficients will remain unchanged in the modelling of a dichotomized response, even when the errors are simulated from a normal distribution. This relationship is exploited in Chapter 5, where a continuous variable is dichotomized to create a binary
response in both a simulation study and the analysis of boar taint data.

2.4 Graphical models

A graph is defined by a set of nodes and a set of edges between nodes. In a graphical model, the nodes correspond to random variables and edges represent statistical relationships. Graphical models allow for a graphical representation of the (conditional) independence relation entailed by the joint distribution over the variables in the graph, and can be particularly useful when working in a high-dimensional space (Koller and Friedman, 2009). Three classes of graphical models include directed, undirected, and mixed models. Directed models, such as Bayesian networks, allow for the modelling of directed relationships between variables. Undirected graphical
models, on the other hand, simply model the associations between pairs of variables conditional on all other variables in the graph. An undirected graphical model consists of a set of nodes and a set of undirected edges that connects pairs of (conditionally) dependent nodes. Figure 2.2 shows an example of an undirected graph consisting of six nodes (labelled 1 through 6) and lines are used to represent the edges, or pairwise associations, between nodes (variables).

Often, the nodes of the graph correspond to the variables measured on an observational unit. Consider an experiment on the nature of the relationships among many genes. As a graphical model, the individual genes would constitute the nodes of the graph while edges between gene pairs would represent a statistical association; the exact association depends on the class of graphical model. It may be found that edges exist between genes involved in the same pathway or alternatively between genes that exhibit significant co-expression. The graphical structure may be defined through scientific knowledge collected a priori or may be estimated from data.

Researchers are often interested in different properties that help to characterize a graph. Two important features are the neighbours and the degree of a node. A neighbour of a node $i$ is any node $j, j \neq i$, such that an edge is present between nodes $i$ and $j$. Such nodes are also called adjacent. The degree of a node $i$, $d_i$, is then the number of nodes to which it is adjacent while the neighbourhood of node $i$, $\mathcal{N}_i$, is the union of node $i$ and the set of all its neighbours. For the graph in Figure 2.2, the neighbours of node 1 are nodes $\{3, 4, 5\}$, $\mathcal{N}_1 = \{1, 3, 4, 5\}$, and its degree is three. In fact all nodes in the graph have degree three.
2.4.1 Gaussian graphical models

Let $Z$ be an $n \times p$ matrix of observations representing measurements of $p$ variables on each of $k = 1, \ldots, n$ individuals. Assume that the $p$-dimensional observations are independently and identically distributed multivariate Gaussian:

$$Z_k \sim \text{Multivariate Gaussian}(\mu, \Sigma),$$

with precision matrix:

$$\Omega = [\omega_{ij}]_{i,j=1}^p = \Sigma^{-1}.$$

Then edges will exist between pairs of nodes, $i$ and $j$, that are dependent conditional on all other nodes in the graph, i.e., $\omega_{ij} \neq 0$. In contrast, for pairs of nodes, $i$ and $j$,
that are independent conditional on all other nodes in the graph, i.e., $\omega_{ij} = 0$, there will be an absence of edges. To illustrate these concepts, suppose the joint distribution of the variables in the graph shown in Figure 2.2 is multivariate Gaussian. As an edge is present between nodes 1 and 3, these nodes are conditionally dependent given the four remaining nodes, \{2, 4, 5, 6\}, and $\omega_{13} = \omega_{31} \neq 0$. However, since there is no edge between nodes 1 and 2, these two nodes are conditionally independent given the remaining four nodes, \{3, 4, 5, 6\}, and $\omega_{12} = \omega_{21} = 0$.

In the case of Gaussian graphical models, the underlying graph structure can be learned through the precision matrix $\Omega$. Since there will be an absence of an edge between pairs of nodes $i, j$ for which $\omega_{ij} = 0$, encouraging sparsity in the precision matrix corresponds to sparsity in the associated graph, which may be advantageous in many scenarios (Friedman et al., 2008). Recall that the $\ell_1$ penalty of the LASSO encourages sparsity in the estimated regression parameter vector by setting some of the coefficient estimates to zero. Similarly, the graphical LASSO applies the $\ell_1$ penalty in the sparse estimation of the precision matrix, $\Omega$, and thereby results in sparsity of the undirected graph (Friedman et al., 2008).

2.5 Regularized regression incorporating graphical structure

In the LASSO model (see Equation (2.5)), the penalty term was calculated as a norm of the regression coefficients weighted by the tuning parameter, $\lambda$. However, this penalty term may be reformulated such that it depends upon the graphical structure of the predictors. Yu and Liu (2016) developed a regularized regression
method that requires a graph of the predictors as input and assumes that this graph is specified correctly.

2.5.1 Sparse regression incorporating graphical structure among predictors: SRIG

In the sparse regression incorporating graphical structure among predictors (SRIG) model (Yu and Liu, 2016), each node contributes to the penalty term such that these contributions depend on the neighbourhood of the node. Let $X$ be an $n \times p$ matrix of predictors and assume that the observations, $X_k, k = 1, \ldots, n$, are independent and identically distributed multivariate Gaussian with mean vector $0_p$, variance-covariance matrix $\Sigma$ and precision matrix $\Omega$. An undirected graph over the variables in $X_k$ would consist of $p$ nodes, say $\{1, 2, \ldots, p\}$, and edges would be present between all pairs of distinct nodes $i, j$ for which the $(i, j)^{th}$ element of $\Omega$ is non-zero, i.e., wherever $\omega_{ij} \neq 0$.

Let $E$ be a $p \times p$ adjacency (edge) matrix where $E_{ij}$ takes on a value of one if an edge is present between nodes $i$ and $j$; 0 otherwise. $E$ may be constructed utilizing pre-existing knowledge or can be estimated using a graphical modelling technique such as the graphical LASSO (Friedman et al., 2008). If $\Sigma_{xy} = (c_1, c_2, \ldots, c_p)^T$ is the vector of covariances between $x_i$ and $y$, $i = 1, \ldots, p$, then under the linear model defined in
Equation (2.1), the regression coefficients are given by \( \beta = \Sigma_{xy} \Omega \), or equivalently,

\[
\begin{align*}
\beta_1 &= c_1 \omega_{11} + c_2 \omega_{12} + \cdots + c_i \omega_{1i} + \cdots + c_p \omega_{1p}, \\
\beta_2 &= c_1 \omega_{21} + c_2 \omega_{22} + \cdots + c_i \omega_{2i} + \cdots + c_p \omega_{2p}, \\
&\vdots \\
\beta_p &= c_1 \omega_{p1} + c_2 \omega_{p2} + \cdots + c_i \omega_{pi} + \cdots + c_p \omega_{pp}. \\
\end{align*}
\] (2.13)

Let \( V^{(i)}_{j} = c_i \omega_{ji} \) for \( i, j = 1, \ldots, p \), represent the contribution of the \( i \)th node to the \( j \)th regression coefficient and \( V^{(i)} = [V^{(i)}_1, \ldots, V^{(i)}_p]^T \) be a \( p \times 1 \) column vector representing the contributions of the \( i \)th predictor to each of the \( p \) regression coefficients for \( i = 1, \ldots, p \). Then, it follows that:

\[
\beta = \sum_{i=1}^p V^{(i)}. \\
\] (2.14)

Since \( \omega_{ij} = 0 \) for pairs of conditionally independent nodes, the support of \( V^{(i)} \) is simply its neighbourhood, \( N_i \). Therefore, node \( i \) will only contribute to the estimation of regression coefficients associated with its neighbours. The \( j \)th element of \( V^{(i)} \) is non-zero provided an edge exists between nodes \( i \) and \( j \) in the predictor graph associated with \( X \). Therefore, the decomposition of \( \beta \) may alternatively be written as a product
of the components of matrix $V = [V^{(1)}, V^{(2)}, \ldots, V^{(p)}]$ and edge matrix $E$:

$$
\beta_1 = V^{(1)}_1 E_{11} + V^{(2)}_1 E_{12} + \cdots + V^{(i)}_1 E_{1i} + \cdots + V^{(p)}_1 E_{1p},
$$

$$
\beta_2 = V^{(1)}_2 E_{21} + V^{(2)}_2 E_{22} + \cdots + V^{(i)}_2 E_{2i} + \cdots + V^{(p)}_2 E_{2p},
$$

$$
\vdots
$$

$$
\beta_p = V^{(1)}_p E_{p1} + V^{(2)}_p E_{p2} + \cdots + V^{(i)}_p E_{pi} + \cdots + V^{(p)}_p E_{pp}. \quad (2.15)
$$

Although both Equations (2.13) and (2.15) represent equivalent decompositions of the regression coefficients, Equation (2.13) highlights the decomposition of the regression coefficients in terms of the precision matrix, while Equation (2.15) expresses the decomposition of the regression coefficients in terms of the graph edges directly.

Finally, the estimates of the regression parameters for the SRIG model are found by solving:

$$
\arg\min_{\beta, V^{(1)}, V^{(2)}, \ldots, V^{(p)}} \frac{1}{2n} ||Y - X\beta||_2^2 + \lambda \sum_{i=1}^{p} \tau_i ||V^{(i)}||_2, \quad (2.16)
$$

where $\tau_i$ is a weight for the $i^{th}$ predictor. In particular, $\tau_i$ is based on the tuning parameter $\gamma > 0$ and is defined as: $\sqrt{\hat\sigma_i} / |\beta_i|$ for $n \geq p$ and $\sqrt{\hat\sigma_i} / |\hat c_i|$ for $n < p$, where $\hat c_i$ is the sample covariance between $Y$ and $X_i$. From Equation (2.16), it can be seen that this penalized regression encourages the contributions of the $i^{th}$ predictor to each of the $p$ regression coefficients to be zero (or non-zero) together as shrinkage and selection occurs on the $V^{(i)}$, $i = 1, \ldots, p$. 

Parameter estimates for the SRIG model may be found using the predictor duplication method as proposed in Obozinski et al. (2011) in conjunction with existing methods for the overlapping group LASSO. This method requires an augmented predictor matrix, \( \tilde{X} \), that is created by replicating the columns of the original predictor matrix once for each neighbourhood in which a given predictor belongs. Alternatively, Yu and Liu (2016) also proposed the iterative proximal (IP) algorithm based on the Fast Iterative Shrinkage Thesholding Algorithm (Beck and Teboulle, 2009) that does not require the duplication of predictors. Consequently, the IP algorithm is more efficient when working in a high dimensional predictor space or when the predictor graph is dense.

### 2.5.2 Limitations of SRIG and thesis overview

Two key issues that arise in practice when trying to incorporate structural information in a regularized regression are: (i) \( X \) may not include all of the relevant predictors of \( Y \); and (ii) the graphical structure of the predictors may be unknown. This latter requirement is important when trying to link the undirected graph over the predictors with the response, \( Y \). Let \( J_0 = \{i : \beta_i \neq 0, i = 1, ..., p\} \) represent the set of nodes that are true predictors of \( Y \), i.e., the set of all nodes for which their associated regression coefficients are non-zero in the underlying data-generating mechanism. Yu and Liu (2016) derived a finite sample error bound for SRIG under certain assumptions, one of which is the following: if the \( i^{\text{th}} \) node is a true predictor of \( Y \) then \( N_i \subseteq J_0 \). This condition states that all neighbours of relevant predictors \( (\forall j \in N_i, j \neq i, i \in J_0) \) are also relevant predictors themselves (nodes \( j, j \neq i \), have
Consider a violation of this assumption in which an edge exists between the pair of nodes \((i, j)\) and \(i \in J_0\) but \(j \in J^C_0\). In some scenarios, it may be reasonable to expect that the associated \(V_j^{(i)}\) will be small relative to the other components in \(V^{(i)}\). Therefore, we may wish to shrink it to zero, while retaining the remaining components.

Further one may question whether it is reasonable to always assume that all the non-zero components within \(V^{(i)}\) be shrunk to zero simultaneously. Perhaps only small elements in \(V^{(i)}\), i.e., individual components \(V_{j}^{(i)}\), \(j = 1, \ldots, p\), should be shrunk to zero while larger elements are retained. To elaborate, note that all elements in \(V^{(i)}\) share a common factor of \(c_i\) (see Equation (2.13)). Therefore, it may be reasonable to wish to shrink contributions where the conditional relationship is relatively weak (small \(\omega_{ji}\)) while retaining those contributions where the conditional relationship is strong (large \(\omega_{ji}\)).

SRIG also implicitly assumes that the true graphical structure of the predictors is known. However, in practice this structure is typically estimated. Whenever a false positive edge is present between nodes \(i\) and \(j\) in the estimated predictor graph, the support of \(V^{(i)}\) may be assumed to be larger than it really is, which may ultimately bias estimates of \(\beta\). So, if we find that \(V_j^{(i)}\) is small relative to the other non-zero components \(V_{j'}^{(i)}, j' \neq j\), we may again wish to shrink it to zero while retaining the other components.

Chapter 3 addresses these limitations and proposes the new DSRIG method where sparsity is encouraged within the \(V^{(i)}, i = 1, \ldots, p\), through the addition of an
ℓ₁ penalty to the SRIG objective function. It is shown that DSRIG has improved predictive performance and decreased bias when compared to the SRIG model, particularly when the predictor graph is unknown and is estimated from the data.

Up to now, it has been assumed that $X_k, k = 1, ..., p$, follow a multivariate normal distribution. However, SNP genotypes are often encoded as $\{0, 1, 2\}$ and therefore will not meet this normality assumption. In Chapter 4, DSRIG is applied in the modeling of the expression of the boar taint compounds androstenone and skatole from a set of candidate SNPs. DSRIG is shown to provide a predictive benefit when compared to ordinary least squares and the LASSO in a cross-validation procedure.

Lastly, SRIG was developed in the context of a quantitative response. However it may be interesting to determine if these methods can be extended to the case of a binary outcome. Chapter 5, extends these methods to a binary outcome in the new DSLRIG model. Through simulation, it was found that DSLRIG offers improved predictive performance to conventional regularized methods as well as an improved ability to both identify and maintain the relative ranking of the true non-zero regression coefficients.
Chapter 3

Doubly Sparse Regression Incorporating Graphical Structure Among Predictors

Matthew Stephenson†, R. Ayesha Ali†, and Gerarda A. Darlington†∗

†Department of Mathematics and Statistics, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

∗Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

3.1 Abstract

Recent research has demonstrated that information learned from building a graphical model on the predictor set of a regularized linear regression model can be leveraged to improve prediction of a continuous outcome. In this paper, we present a new model that encourages sparsity at both the level of the regression coefficients and at the level of individual contributions in a decomposed representation. This model provides parameter estimates with a finite sample error bound and exhibits

---

Chapter 3 is under revision for publication in The Canadian Journal of Statistics. The supplementary information referred to in Chapter 3 can be found in Appendix A.
robustness to errors in the input graph structure. Through a simulation study and
the analysis of two real world data sets, we demonstrate that our model provides a
predictive benefit when compared to previously proposed models. Further, it is a
highly flexible model that provides a unified framework for the fitting of many other
commonly used regularized regression models.

3.2 Introduction

A common objective in statistical learning is the prediction of an outcome
$Y$ from a set of input features $X$. For example, how can volume measurements
of various brain regions be used in the prediction of cognitive impairment? Linear
regression has proven to be a simple yet effective technique that can often outperform
more complicated non-linear approaches, particularly in cases of low signal-to-noise
ratio, small training sets or sparse data (Hastie et al., 2001). However, as data sets
continue to increase in size, the use of conventional statistical methods is challenged
when the number of input features (explanatory variables), $p$, exceeds the sample size,
$n$. Therefore, in addition to accurate prediction, a secondary (and often inherent)
objective of many predictive models involves variable selection. In the context of
a brain region volume example, we may be interested in not only the prediction
of cognitive impairment, but also the identification of regions of the brain that are
associated with cognitive functioning.

Regularized regression adds a penalty term $\mathcal{R}(\beta)$, weighted by a tuning
parameter $\lambda > 0$, to an ordinary least squares model and has been proposed as one
method to improve predictive accuracy. These models provide parameter estimates
even when the matrix of predictors, $X$, is not of full rank and are valid in the presence
of multicollinearity or when $p > n$ (Hastie et al., 2001). More recently Yu and Liu (2016) introduced structural sparsity among the regression coefficients by assuming a graphical structure of the predictors and then exploiting that structure during minimization.

Here, we introduce the Doubly Sparse Regression Incorporating Graphical structure among predictors (DSRIG) model. This regularized regression model provides parameter estimates with finite sample error bound and is more robust to the presence of false-positive edges in the predictor graph. The graphical structure of the predictors is first modeled independently from the outcome and this predictor structure is then leveraged to improve prediction in the construction of $\mathcal{R}(\beta)$. DSRIG allows for sparsity both at the level of contributions of the regression coefficients to the outcome but also at the level of individual contributions of the predictors to $\beta$ based on a decomposed representation of the regression coefficients. Our new model has improved prediction, is highly flexible, and provides a unified framework for fitting many other regularized linear regression models.

Section 3.3 motivates and introduces our new model, DSRIG, and outlines the estimation procedure. Section 3.4 theoretically evaluates parameter estimates and derives the finite sample error bound. Section 3.5 examines the empirical properties of the estimates through a comprehensive simulation study, while Section 3.6 presents the analysis of two real world data sets. Lastly, Section 3.7 discusses our results, provides conclusions and highlights areas for future research.
3.3 Methods and Motivation

Let $\mathcal{J}$ represent a set of nodes defined by $p$ predictors labeled $1, \ldots, p$, each measured on $n$ individuals such that the predictor matrix $\mathbf{X}$ can be arranged in an $n \times p$ matrix of observations. Assume that the $p$-dimensional observations on each individual are independently and identically distributed multivariate normal (MVN), $\mathbf{X}_k \sim \text{MVN}(\mathbf{\mu}, \mathbf{\Sigma}), k = 1, \ldots, n$, with $p \times p$ precision matrix $\mathbf{\Omega}$ taking elements $\omega_{ij}$ for $i, j = 1, \ldots, p$; i.e., $\mathbf{\Omega} = [\omega_{ij}]_{i,j=1}^{p} = \mathbf{\Sigma}^{-1}$. Then, an undirected graph representing the joint distribution of the variables in $\mathbf{X}$ will have edges between any nodes $(i, j)$, $i \neq j$, wherever $\omega_{ij} \neq 0$.

Let $\mathbf{\Sigma}_{xy} = (c_1, c_2, \ldots, c_p)^T$ be the cross covariance vector between predictors in $\mathbf{X}$ and an $n \times 1$ response vector $\mathbf{Y}$. Then, under the linear model $\mathbf{Y} = \mathbf{X}\mathbf{\beta} + \mathbf{e}, \mathbf{e} \sim \text{MVN}(0, \sigma^2\mathbf{I}_n)$, the $p \times 1$ regression coefficient vector $\mathbf{\beta}$ may be decomposed as $\mathbf{\beta} = \mathbf{\Omega}\mathbf{\Sigma}_{xy}$ i.e.,

\[
\beta_1 = c_1\omega_{11} + c_2\omega_{12} + \cdots + c_i\omega_{1i} + \cdots + c_p\omega_{1p}, \\
\beta_2 = c_1\omega_{21} + c_2\omega_{22} + \cdots + c_i\omega_{2i} + \cdots + c_p\omega_{2p}, \\
\vdots \\
\beta_p = c_1\omega_{p1} + c_2\omega_{p2} + \cdots + c_i\omega_{pi} + \cdots + c_p\omega_{pp}. \quad (3.1)
\]

Let $V_j^{(i)} = c_i\omega_{ij}$ for $i, j = 1, \ldots, p$, represent the contribution of the $i^{th}$ predictor to the $j^{th}$ regression coefficient and let $\mathbf{V}^{(i)} = [V_1^{(i)}, \ldots, V_p^{(i)}]^T$ be a $p \times 1$ column vector representing the contributions of the $i^{th}$ predictor to $\mathbf{\beta}$. Then, the equations in (3.1) can be re-expressed as $\mathbf{\beta} = \sum_{i=1}^{p} \mathbf{V}^{(i)}$. Further let the neighbourhood of node $i$, $\mathcal{N}_i$, be defined as the union of node $i$ and the set of all its neighbours, i.e., the set
of all other nodes $j = 1, \ldots, p, j \neq i$, such that $\omega_{ij} \neq 0$. Consequently, the support of $V^{(i)}$ will simply be given by the neighbourhood $N_i$ since $\omega_{ji} = 0$ whenever there is no edge between nodes $(i, j)$ in the predictor graph. In other words, node $i$ only contributes to the estimation of regression coefficients associated with its neighbours and so learning the support of $V^{(i)}$ is analogous to learning the structure of the predictor graph. Yu and Liu (2016) exploit this relation in the SRIG model.

### 3.3.1 SRIG

The SRIG model of Yu and Liu (2016) assumes that the predictor graph structure is known. As such, the support of $V^{(i)}$ is also known and estimation of the SRIG coefficients proceeds by solving:

$$
\arg\min_{\beta, V^{(i)} \forall i} \left\{ \frac{1}{2n} \| Y - X\beta \|_2^2 + \lambda \sum_{i=1}^{p} \tau_i \| V^{(i)} \|_2 \right\}, \text{ where } \beta = \sum_{i=1}^{p} V^{(i)},
$$

(3.2)

where $\lambda \geq 0$ is a tuning parameter and $\tau_i$ is a weight for the $i^{th}$ predictor. In particular, $\tau_i$ is based on another tuning parameter $\gamma > 0$ and $\tau_i = \frac{\sqrt{d}}{|N_i|^{\gamma}}$ for $n \geq p$ and $\tau_i = \frac{\sqrt{d}}{|N_i|^{\gamma}}$ for $n < p$, where $d_i = |N_i|$ is the degree of node $i$ and represents the size of its neighbourhood. Although the tuning parameter $\gamma$ may be trained via cross-validation we will assume $\gamma = 1$ from here forward. Note that the SRIG penalty in Equation (3.2) is similar to the group-LASSO penalty of Yuan and Lin (2006) but acts only on the vectors of contributions $\{V^{(i)}, i = 1, \ldots, p\}$. That is, the shrinkage and selection of the SRIG model is done at the level of the $V^{(i)}$'s rather than directly on the regression coefficients in $\beta$. However, there may be situations in which encouraging sparsity within the $V^{(i)}$'s is desirable.

Consider the augmented graph consisting of the union of the predictors in $X
and the outcome $Y$. Let $\mathcal{J}_0 = \{ i : \beta_i \neq 0, i = 1, \ldots, p \}$ represent the set of nodes that are true (direct) predictors of $Y$, i.e., the neighbours of $Y$ in the augmented graph (e.g., $X_1$ and $X_2$ of Figure 3.1(a)). Equivalently, the set of true (direct) predictors can be thought of as the set of nodes associated with non-zero regression coefficients in the underlying data-generating mechanism. Yu and Liu (2016) demonstrated that the finite sample bounds derived for SRIG prediction and estimation errors require that if any node $i \in \mathcal{J}_0$, then $\mathcal{N}_i \subseteq \mathcal{J}_0$ (Assumption A2, Section 4.3 of Yu and Liu (2016)). In other words, all neighbours of true predictors are also true predictors. This implies that any two nodes connected by a continuous path in the predictor graph are assumed to have associated regression coefficients that are zero or non-zero together. We too will make use of this assumption (see Lemma 3.3).

Suppose edge $(i, j)$ is in the predictor graph such that node $i \in \mathcal{J}_0$ but node $j \notin \mathcal{J}_0$. For example, consider the edge between $X_2$ and $X_3$ of Figure 3.1(b). SRIG would set $V^{(3)}_1 = V^{(1)}_3 = 0$ since $\omega_{13} = \omega_{31} = 0$. However, the vector of contributions $V^{(3)}$ associated with $X_3$ may not be shrunk to zero, thereby increasing the bias in $\beta$. Further, SRIG does not provide a mechanism to shrink $V^{(2)}_3$ to 0 without (erroneously) shrinking the entire vector $V^{(2)}$ to 0. Therefore, SRIG may produce a non-zero $\beta_3$, which would be equivalent to adding an edge between $X_3$ and $Y$ in the augmented graph of Figure 3.1(a). Through simulation, Yu and Liu (2016) found that SRIG
performance decreased relative to the LASSO as the assumption that for any node $i \in J_0$, then $N_i \subseteq J_0$ became increasingly challenged. By allowing for the shrinkage of small $V_j^{(i)}$ associated with an edge $(i, j)$ for which $i \in J_0$ but $j \notin J_0$ to zero the effects of violations to this assumption can be minimized.

From another perspective, in practice the predictor graph structure is typically unknown and estimated from data. Accordingly, $(i, j)$ may be a false positive edge in the predictor graph with a weak $V_j^{(i)}$ relative to $V_k^{(i)}$ for some $k \in N_i, k \neq j$. Shrinking such $V_j^{(i)}$ to zero would make the estimation procedure more robust to predictor graph misspecification and help mitigate bias in the final estimate of $\beta$. In short, a model that further encourages sparsity within the $V^{(i)}$ by shrinking small $V_j^{(i)}$ to zero while retaining the other larger components can improve variable selection and predictive performance.

### 3.3.2 DSRIG

In order to induce sparsity both between and within the $V^{(i)}, i = 1, \ldots, p$, the new DSRIG model adds an $\ell_1$ penalty to the SRIG objective function with regression coefficients found as the solution to:

$$
\arg\min_{\beta, V^{(i)}_V} \left( \frac{1}{2n} \| Y - X\beta \|_2^2 + \lambda \left\{ \sum_{i=1}^p [\tau_i \| V^{(i)} \|_2 + \xi \| V^{(i)} \|_1] \right\} \right), \beta = \sum_{i=1}^p V^{(i)}, \quad (3.3)
$$

for $\lambda \geq 0$ and $\xi \geq 0$. The tuning parameter $\xi$ balances the contributions of the $\ell_1$ and $\ell_2$ components of the penalty term. The penalty proposed in Equation (3.3) is similar in nature to the sparse-group-LASSO (Simon et al., 2013) but performs shrinkage on the $V^{(i)}$’s rather than directly on $\beta$. The $\ell_1$ component induces sparsity within $V^{(i)}$ by shrinking individual contributions $V_j^{(i)}$ to zero, while the $\ell_2$ penalty functions as
in SRIG and shrinks entire vectors of contributions, **V**\(^{(i)}\), to the zero vector.

As with SRIG, in practice DSRIG typically first requires estimation of the predictor graph. To further mitigate any potential bias induced by graph misspecification, we recommend that the predictor graph structure estimation be incorporated in a cross-validation scheme in which the graph structure is learned on only the training set. The optimization problem in Equation (3.3) can then be solved across a grid of \((\lambda, \xi)\) using the training data with the optimal tuning parameters being chosen by an independent validation set. Since the learned predictor graphs from the training sets would exhibit variation, the estimated model parameters would implicitly reflect some of this uncertainty in graph structure.

### 3.3.3 Estimation by Proximal Gradient Descent

The DSRIG regression parameters can be estimated using proximal gradient descent. First reformulate the DSRIG optimization problem (Equation (3.3)) such that the predictor matrix **X** is represented in expanded form (Obozinski et al., 2011; Rao et al., 2013, 2014; Yu and Liu, 2016). Let **X**\(^{(j)}\) represent the \(j^{th}\) column of the original predictor matrix **X**, the \(n \times 1\) vector of measurements for predictor \(j\). Further let \(\tilde{\mathbf{X}} = [\mathbf{X}^{(j)}], j \in \mathcal{N}_i, i = 1, \ldots, p\), be an augmented predictor matrix with columns consisting of replicates of the columns of the original predictor matrix, **X**\(^{(j)}\), for each neighbourhood in which predictor \(j\) belongs; therefore, \(\tilde{\mathbf{X}}\) will be of dimension \(n \times \sum d_i\). When \(\mathcal{N}_i = \{i\} \forall i\) (no edges in the predictor graph), then \(\tilde{\mathbf{X}} = \mathbf{X}\). If \(\tilde{\mathbf{V}}\) is defined as the \(\sum d_i \times 1\) column vector formed by concatenating all the non-zero elements of the \(\mathbf{V}^{(i)}, i = 1, \ldots, p\), then the expected value of **Y** can be found as: \(\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta = \tilde{\mathbf{X}}\tilde{\mathbf{V}}\), and Equation (3.3) in its re-expressed expanded
form is given by:

\[
\arg\min_{\mathbf{V}^{(i)\cap i}} \left( \frac{1}{2n} \| \mathbf{Y} - \mathbf{X} \mathbf{V} \|^2 + \lambda \left( \sum_{i=1}^{p} \tau_i \| \widetilde{\mathbf{V}}^{(i)} \|_2 + \xi \| \mathbf{V} \|_1 \right) \right)
\]

\[
= \arg\min_{\mathbf{V}^{(i)\cap i}} \left( \mathcal{L}(\beta) + \lambda \mathcal{R}(\beta) \right), \quad \beta = \sum_{i=1}^{p} \mathbf{V}^{(i)}
\]  

(3.4)

where \( \widetilde{V}^{(i)} \) represents the non-zero components in \( \mathbf{V}^{(i)} \), \( \mathcal{L}(\beta) \) is a smooth loss function and \( \mathcal{R}(\beta) \) is a non-smooth function (penalty term).

Note that, by definition, \( \widetilde{V}^{(i)} \) contains the contributions from the neighbourhood of \( i \), \( \mathcal{N}_i \). In fact, the predictor graph can be viewed as a set of neighbourhoods that may overlap. Consequently, the optimization in Equation (3.4) is analogous to the multi-task set up in Rao et al. (2013) in which the neighbourhoods represent a set of overlapping groups and the optimization induces sparsity both among and within groups (neighbourhoods). Accordingly, we can utilize proximal point methods that alternate between taking a gradient step in the negative direction of \( \mathcal{L}(\beta) \) followed by subsequent application of the proximal operator of \( \mathcal{R}(\beta) \) to do the optimization (Rao et al., 2014).

The gradient of \( \mathcal{L}(\beta) \) with respect to \( \mathbf{V} \) is: \( \nabla \mathcal{L}(\beta) = \frac{1}{n} \mathbf{X}' \left( \mathbf{X} \mathbf{V} - \mathbf{Y} \right) \). So, the parameter estimates for iteration \( r + 1 \) can be found as:

\[
\mathbf{V}^{(r+1)} = \text{Prox} \left[ \mathbf{V}^{(r)} - t \nabla \mathcal{L} \left( \mathbf{V}^{(r)} \right) \right],
\]

for some step size \( t \), where \( \text{Prox}[-] \) is the proximal operator of \( \mathcal{R}(\beta) \). Let \( \mathbf{V}_\nabla \) be the intermediary found after taking a gradient step, but before applying the proximal operator. This proximal operator can be divided into two steps: (i) a soft thresholding
of the individual elements $\tilde{V}_j, j = 1, \ldots, \sum_{i=1}^p d_i$, to address the $\ell_1$ component:

$$\tilde{V}^*_j = \begin{cases} 
\text{sign}(\tilde{V}_{\nabla j}) - \lambda \xi & \text{for } |\tilde{V}_{\nabla j}| > \lambda \xi \\
0 & \text{otherwise}
\end{cases}; \quad (3.5)$$

and (ii) a group soft thresholding of each of the $\tilde{V}^{(i)}, i = 1, \ldots, p$, to address the $\ell_2$ component:

$$(\tilde{V}^{(i)})^{\{r+1\}} = \begin{cases} 
\frac{\tilde{V}^{(i)*}}{\|\tilde{V}^{(i)*}\|_2} \left(\|\tilde{V}^{(i)*}\|_2 - \lambda \tau_i\right) & \text{for } \|\tilde{V}^{(i)*}\|_2 > \lambda \tau_i \\
0 & \text{otherwise}
\end{cases}. \quad (3.6)$$

From the proximal operator steps in Equations (3.5) and (3.6), it is easy to see the hierarchical structure of our proximal operator where shrinkage and selection is first performed element-wise ($\tilde{V}_j$) and then performed group-wise ($\tilde{V}^{(i)}$). In practice, we actually implemented the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) with backtracking to choose the step size $t$ (Beck and Teboulle, 2009) in order to accelerate convergence.

### 3.4 Estimator Properties

Since DSRIG is designed to accommodate the $p > n$ problem, standard consistency of the DSRIG estimator cannot be obtained. Instead, we derive an explicit finite sample error bound that holds with high probability. To find such a bound is not straightforward. First, note that the derivation of the consistency rate in Yu and Liu (2016) is not easily extended to the case where $\mathcal{R}(\beta)$ includes an $\ell_1$-penalty term.
Second, while we may borrow ideas from Rao et al. (2013), our bound makes different assumptions on the tuning parameters and is derived using the notation introduced for the undirected predictor graph. Specifically, in their derivation of a finite sample error bound, Rao et al. (2013) required equality of the group weights, $\tau_i, i = 1, \ldots, p$; and an equal contribution of the $\ell_1$ and $\ell_2$ penalties to $R(\beta)$, i.e., $\tau_i = \xi = 1 \forall i$. However, here we derive a bound without requiring these assumptions.

Negahban et al. (2012) provide a framework for deriving finite sample error bounds for convex optimization problems like that in Equation (3.3) when $\lambda > 0$ and $R(\beta)$ is a norm. Fortunately, several of the results needed here fall out as corollaries to results proven in Negahban et al. (2012); for self-containment purposes, we will provide these results without proof, as needed. This framework hinges on two key properties of the objective function to be minimized: (a) the regularizer, or penalty function, $R(\beta)$ is decomposable; and (b) the loss function $L(\beta)$ meets a restricted strong convexity condition. The following assumptions will be needed to establish these two properties.

(A1) The decomposition of our regression coefficients into the set of vectors $V^{(i)}, i = 1, \ldots, p$, is an optimal decomposition;

(A2) For any node $i \in J_0$, then $N_i \subseteq J_0$;

(A3) The true regression parameter vector $\beta$ is exactly sparse with $s$ non-zero components that can be decomposed into a set of $a$ active vectors $V^{(i)}$ with at most $d^{max} = \max_{i=1,\ldots,p} (d_i)$ non-zero elements;

(A4) $\tau^{max} = \max(\tau_i) \forall i$, is upper bounded by $\xi$, and $\tau^{min} = \min(\tau_i) \forall i$, is lower bounded by 1;
(A5) The loss function $L(\beta)$ satisfies the restricted strong convexity (RSC) conditions with curvature parameter $\kappa_L$ (see Definition 3.5);

(A6) The design matrix $X$ is fixed; the observation errors $\epsilon_k, k = 1, \ldots, n$, are additive, independent of $X$ and $\epsilon_k \overset{i.i.d.}{\sim} \text{Normal } (0, \sigma)$.

Based on these assumptions, we can obtain a finite sample error bound for the $\ell_2$ error between any optimal solution, $\hat{\beta}$, of Equation (3.3) and the true parameter vector $\beta$. Assumption (A4) explicitly states $\tau_{\min} \geq 1$; however, $X$ and $Y$ are typically scaled to have columnwise standard deviations of 1 and therefore, $|c_i| \leq 1$, $d_i \geq 1$ and for $\tau_i = \frac{\sqrt{d_i}}{|c_i|}$ we have $\tau_{\min} \geq 1$. Although (A4) is needed to obtain the theoretical results presented, we found that prediction improved when $\max(\tau_i) \leq \xi, i = 1, \ldots, p$, is not enforced and all analysis results reported here are with this restriction removed.

We now state our main result.

**Theorem 3.1.** Assume (A1)-(A6) for the optimization problem in Equation (3.3) and define $\sigma_i^*$ to be the maximum singular value of $X_i^T X_i$ and $\sigma_{\max} = \max_{i=1,\ldots,p} (\sigma_i^*)$. Then for $\lambda^2 \geq \sigma^2 \sigma_{\max} (\log(p) + d_{\max})$ \((\tau_{\min})^2\), any optimal solution, $\hat{\beta}$, will satisfy:

$$
\|\hat{\beta} - \beta\|_2^2 \leq \frac{9 \sigma^2 \sigma_{\max} (\tau_{\max} + \xi \sqrt{d_{\max}})^2 a(\log(p) + d_{\max})}{n \kappa_L},
$$

with probability at least $1 - c_1 \exp(-c_2 \sqrt{n})$ for some $c_1, c_2 > 0$.

Section 3.4.1 provides definitions of decomposability and restricted strong convexity, and proves that the DSRIG optimization problem in Equation (3.3) exhibits these properties. Section 3.4.2 derives the finite sample error bound presented in Equation (3.7). Proofs and intermediary results are contained in Appendix A.
3.4.1 Properties of $\mathcal{R}(\beta)$ and $\mathcal{L}(\beta)$

Assume (A2), that for any node $i \in \mathcal{J}_0$ we have $\mathcal{N}_i \subseteq \mathcal{J}_0$. Let $\mathcal{M}$ and $\overline{\mathcal{M}}$ be two subspaces such that $\mathcal{M} \subseteq \overline{\mathcal{M}} \in \mathbb{R}^p$ and let $\overline{\mathcal{M}}^\perp$ be the orthogonal complement of $\overline{\mathcal{M}}$. In what follows, we choose $\mathcal{M}$ to be the model subspace, which reflects the constraints imposed by DSRIG. Define the cardinality of the set of true predictors $\mathcal{J}_0$, $|\mathcal{J}_0| = s$ such that $s << p$ and assume $\beta$ is exactly sparse, i.e., $\beta \in \mathcal{M}$ where $\mathcal{M}$ can now be further defined as the $s$-dimensional model subspace spanned by the coordinates indexed by $\mathcal{J}_0$. Then $\overline{\mathcal{M}}^\perp = \mathcal{M}^\perp$ may be defined as the subspace spanned by the remaining $p - s$ coordinates indexed in $\mathcal{J}_0^c$. Similarly, define $A_0 = \{i : V^{(i)} \neq 0, i = 1, \ldots, p\}$, with cardinality $|A_0| = a$, where $a << p$, to be the set of active $V^{(i)}$ in our decomposition of $\beta$. If $d_{\text{max}}$ is defined to be the maximum degree across all nodes, then the vectors $V^{(i)}$ themselves are sparse with at most $d_{\text{max}}$ non-zero elements, as stated in assumption (A3). Further note that $\mathcal{J}_0$ must lie within the union of the supports across all active $V^{(i)}$, i.e., $\mathcal{J}_0 = \bigcup_{i \in A_0} \text{supp}(V^{(i)})$. These assumptions will be used in many of our subsequent results.

**Definition 3.2.** (Negahban et al. (2012)) Given the pair of subspaces $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, a norm-based regularizer $\mathcal{R}(\cdot)$ is decomposable if $\mathcal{R}(\beta + \beta^*) = \mathcal{R}(\beta) + \mathcal{R}(\beta^*)$ for all $\beta \in \mathcal{M}$ and $\beta^* \in \overline{\mathcal{M}}^\perp$.

**Lemma 3.3.** Assume (A1)-(A3). Then $\mathcal{R}(\beta)$ is a norm and decomposable with respect to the subspace pair $(\mathcal{M}, \mathcal{M}^\perp)$.

Similar to previous work on decomposed regression coefficients (Obozinski et al., 2011; Rao et al., 2013, 2014; Yu and Liu, 2016) we require that our decomposition is *optimal* in the sense that there is no other decomposition for which the associated penalty $\mathcal{R}(\beta)$ is smaller (See Definition S.1 in Appendix A). Optimality of the decomposition
is needed to show that $\mathcal{R}(\beta)$ is a decomposable norm. Rao et al. (2013) note that if $\mathcal{R}(\beta)$ is convex and coercive, which is also the case here, an optimal decomposition exists $\forall \beta$.

We now concentrate on properties of the loss function $\mathcal{L}(\beta)$ and specify the RSC condition needed to establish a finite sample bound. These conditions ensure that there is sufficient curvature in $\mathcal{L}(\beta)$ around its optimum to allow for parameter estimation. In particular, consider the error term in a first-order Taylor series expansion of $\mathcal{L}(\beta)$ in some direction $\Delta$, $\delta\mathcal{L}(\Delta, \beta)$. Since $p > n$, it suffices to show that $\delta\mathcal{L}(\Delta, \beta)$ is lower bounded by $\kappa_\mathcal{L}\|\Delta\|^2$, for some $\kappa_\mathcal{L} > 0$ for all $\Delta$ in a restricted direction. The estimation error, $\hat{\Delta} = \hat{\beta} - \beta$, is the appropriate direction of interest here. Lemma 3.4 shows that $\hat{\Delta}$ falls in a cone set for regularizers with dual norms that bound $\lambda$. This cone set is the appropriate space in which strong convexity is needed. Definition 3.5 states the restricted convexity condition.

**Lemma 3.4.** Suppose $\mathcal{L}(\cdot)$ is a convex and differentiable loss function and consider any optimal solution $\hat{\beta}$ to the optimization problem in Equation (3.3) with a strictly positive regularization parameter satisfying $\lambda \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\beta))$ where $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$ and $\nabla \mathcal{L}(\beta)$ is the gradient of the loss function. Assume (A1)-(A3), and let $\Pi_\mathcal{M}(\cdot)$ represent the projection onto the subspace $\mathcal{M}$. Then, the error, $\hat{\Delta} = \hat{\beta} - \beta$, will belong to the set:

$$
\mathcal{C}(\mathcal{M}, \mathcal{M}^\perp, \beta) := \{\Delta \in \mathbb{R}^p | \mathcal{R}(\Pi_{\mathcal{M}^\perp} \Delta) \leq 3\mathcal{R}[\Pi_\mathcal{M}(\Delta)]\}.
$$

**Definition 3.5.** The loss function $\mathcal{L}(\beta)$ will satisfy an RSC condition with curvature
parameter $\kappa_L$ if it is convex, differentiable, and:

$$\delta L(\Delta, \beta) := L(\beta + \Delta) - L(\beta) - \langle \nabla L(\beta), \Delta \rangle \geq \kappa_L \|\Delta\|^2, \quad \forall \Delta \in C(M, M^\perp, \beta),$$

where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors and $C(\cdot)$ is as defined in Equation (3.8).

Assumption (A5) simply states that the loss function meets this RSC condition. Note that Negahban et al. (2012) provide an RSC condition that involves a tolerance parameter, but since $\beta \in M$, this tolerance parameter is equal to zero and the corresponding term in their lower bound can be ignored.

### 3.4.2 Finite Sample Error Upper Bound

Negahban et al. (2012) provided the following error bound for general models, when Lemma 3.3 and Definition 3.5 hold.

**Theorem 3.6.** (Negahban et al. (2012)) Assume (i) the regularizer $R(\cdot)$ is a norm and decomposable with respect to the subspace pair $(M, M^\perp)$; (ii) the loss function $L(\cdot)$ is convex and differentiable, and satisfies RSC with curvature $\kappa_L$; and (iii) $\beta \in M$. Then, for $\lambda \geq 2R^*(\nabla L(\beta))$ any optimal solution $\hat{\beta}$ to the optimization problem in Equation (3.3) will have error:

$$\|\hat{\beta} - \beta\|^2 \leq \frac{9}{\kappa_L^2} \psi^2(M),$$

where $\psi(\cdot)$ is a subspace compatibility constant that measures the compatibility between the regularizer $R(\cdot)$ and the error norm $\|\cdot\|$. 
Our RSC condition is needed for a regularizer $R(\cdot)$ that is not too large relative to the error norm. The subspace compatibility constant formalizes this notion by explicitly relating the error norm and the regularizer (see Definition S.2 in Appendix A). Establishing a concrete error upper bound requires bounding the subspace compatibility constant $\psi(\cdot)$, per Lemma 3.7, and bounding the dual norm $R^*(\nabla L(\beta))$, per Lemma 3.8 which in turn will provide a more concrete bound on $\lambda$.

**Lemma 3.7.** The subspace compatibility constant associated with the optimization in Equation (3.3) is bounded by:

$$
\psi(\mathcal{M}) \leq \left( \tau^{\text{max}} + \xi \sqrt{d^{\text{max}}} \right) \sqrt{a}.
$$

**Lemma 3.8.** Assume (A4) and (A6). Then:

$$
R^* (\nabla L(\beta))^2 \leq \frac{\sigma^2 \sigma^{\text{max}} (\log(p) + d^{\text{max}})}{4 (\tau^{\text{min}})^2 n},
$$

with probability at least $1 - c_1 \exp(-c_2 \sqrt{n})$ for some $c_1, c_2 > 0$.

Finally, the result in Equation (3.7) can be obtained if we assume, as in Theorem 3.6, that $\lambda \geq 2R^*(\nabla L(\beta))$. Then, from Lemma 3.8, $\lambda^2 \geq \frac{\sigma^2 \sigma^{\text{max}} (\log(p) + d^{\text{max}})}{(\tau^{\text{min}})^2 n}$ which may subsequently be substituted into Equation (3.9).

### 3.5 Simulation Study

Through a comprehensive simulation study we investigate the empirical properties of the DSRIG estimates. Not only does DSRIG have a predictive benefit and decreased bias when compared to SRIG, but results demonstrate that DSRIG also
exhibits robustness to both errors in graph estimation and violations of assumptions.

3.5.1 Simulation Study Design

Our simulation study considered two predictor graph structures: (i) random (sparse); and (ii) scale-free. A scale-free structure comprises relatively few hub-nodes of high degree that connect the rest of the lesser connected nodes in the graph (Figure 3.2), as is commonly seen in biological network models (Jeong et al., 2000). For each predictor graph structure, we generated the precision matrix ($\Omega$) for 30 parent graphs, each with $p = 100$ nodes as follows. Random parent graphs were generated following Yu and Liu (2016) whereby $\Omega = B + \delta I_p$. The diagonal entries of $B$ were initialized to zero while the off diagonal entries of $B$ took on a value of 0.5 with probability 0.05 and 0 with probability 0.95; therefore the probability of an edge between any pair of nodes $(i, j)$ was 0.05. The value of $\delta$ was chosen such that the condition number of $\Omega$ was equal to $p$ and $\Omega$ was subsequently standardized to have unit diagonals. Scale-free parent graphs were generated by first generating a data set with 1,000,000 observations from a scale-free graph using the huge.generator() function of the huge package (Zhao et al., 2015) for R (R Core Team, 2016). The empirical precision matrix, $\Omega$, of these data was then taken to be the true parent precision matrix. From each parent graph, 100 independent $n \times p$ predictor sets, $X$, were generated (see Figure 3.2) using the rmvnorm function of the mvtnorm package (Genz et al., 2017; Genz and Bretz, 2009) for R (R Core Team, 2016).

Nodes were sorted according to their respective degrees. Then, for the random graph scenario, $c_i = 4$ for those nodes $i$ that corresponded to the nodes with the four largest degrees. For the scale-free graph, instead $c_i = 4$ for those nodes $i$ that corresponded to nodes with the second through fifth largest degrees. The value of $c_i$
Figure 3.2: Estimated predictor graphs from independent data sets generated from the same scale-free parent graph.

for all other nodes was assigned to be 0. We considered two scenarios for generating the true regression parameter $\beta$.

In Scenario 1, we wanted to ensure that, for the most part, neighbours of covariates correlated with $Y$ were also correlated with $Y$. We simply set $\beta = \Omega \Sigma_{xy}$. While this calculation does not ensure assumption (A2) is met, the number of edges that violate this assumption is minimized. To challenge (A2), in Scenario 2 we took half of the original non-zero $\beta$'s and switched them with $\beta$'s that were originally calculated as zero. Covariate data were generated for $n = 480$ and $n = 560$ subjects, for a total of $30 \times 100 = 3000$ independent data sets for each predictor graph structure and sample size combination. The response variable, $Y$, was then calculated from the classical linear model, $Y = X\beta + \epsilon, \epsilon \sim \text{MVN}(0, \sigma^2I_n)$, with $\sigma = 5$.

Each data set was partitioned into a training set to learn the model, a validation set to choose the optimal value of the tuning parameter(s) and a test set to
compare the predictive accuracy among the candidate models. For \( n = 480 \), the data were partitioned into train/validation/test set sample sizes of 40/40/400, while for \( n = 560 \), the sample sizes were 80/80/400. For each data set the LASSO (Tibshirani, 1996), SRIG (Yu and Liu, 2016), and DSRIG models were fit and compared. Further, the SRIG and DSRIG models were fit under the best case scenario, which used the true graphical structure (SRIG-True, DSRIG-True) and the more realistic scenario in which the graphical structure was estimated from only the training data (SRIG-Est, DSRIG-Est).

For each data set, the predictor graph for the SRIG and DSRIG models was estimated using the \texttt{huge} and \texttt{huge.select} functions of the \texttt{huge} package (Zhao et al., 2015) in R (R Core Team, 2016), which are based on the Meinshausen-Bühlmann (MB) (Meinshausen and Bühlmann, 2006) method with the STability Approach to Regularization Selection (STARS) criterion (Liu et al., 2010). The optimal value of the tuning parameter \( \lambda \) was chosen from a grid of 100 equally spaced values on a logarithmic scale. The maximum, \( \lambda_{\text{max}} \), was chosen such that it was the minimum value for which \( \hat{\beta} = 0 \) when \( \xi = 0 \), while the minimum was set at \( \lambda_{\text{min}} = 0.01 \times \lambda_{\text{max}} \). Similarly, the optimal \( \xi \) was chosen out of a grid of 100 equally spaced values, but between 0 and \( \xi_{\text{max}} \), where \( \xi_{\text{max}} \) was chosen to be just large enough such that the upper bound was never selected. Lastly, \( \tau_i \) was calculated using the sample covariance with \( \gamma = 1 \).

For each model the final estimated regression parameter vector, \( \hat{\beta} \), was chosen to be that which minimized the prediction error of the validation set (i.e., minimized the Euclidean distance between the observed \( \mathbf{Y} \) and predicted values \( \hat{\mathbf{Y}} \)). The \( \ell_2 \)-distance between the estimated and true regression parameter vectors (\( \|\hat{\beta} - \beta\|_2 \))
Table 3.1: The $\ell_2$-distance and relative prediction error of model fits under two different cross-validation splits for training/validation/test for random and scale-free graphs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40/40/400</td>
<td>80/80/400</td>
</tr>
<tr>
<td>Random Graph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale-Free Graph</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

was used to compare the bias between the candidate models. The relative prediction error (RPE), calculated as: 

$$RPE = \frac{1}{\sigma^2 N_{test}} \left( \hat{\beta} - \beta \right)^T \left( X_{test}^T X_{test} \right) \left( \hat{\beta} - \beta \right),$$

where $X_{test}$ is the predictor matrix for the test set and $N_{test}$ is the size of the test set, was used to compare predictive accuracy. For the case where the predictor graphs were estimated, we also recorded the proportion of times for which DSRIG was preferred, SRIG was preferred or for which the models were equivalent (i.e., DSRIG parameter $\xi = 0$). The absolute median difference of both $\ell_2$-distance and RPE between the two models was also recorded. The proximal gradient methods to estimate the SRIG and DSRIG models were implemented in MATLAB version 2016b (MATLAB, 2016) using code adapted from Cox (2014), while the LASSO was fit using the MATLAB function in the Statistics Toolbox.

### 3.5.2 Simulation Study Results

Table 3.1 records the $\ell_2$-distance between the estimated and true regression parameters as well as the RPE for each of the candidate models. As expected, both
DSRIG and SRIG fit under the true graphical structure have decreased both bias and RPE when compared to the corresponding models fit under an estimated predictor graph. However, DSRIG typically offers improved performance over SRIG both in terms of predictive accuracy and bias, particularly in the estimated predictor graph setting. DSRIG outperformed the LASSO in Scenario 1 while remaining competitive to the LASSO in Scenario 2. However, while SRIG outperformed the LASSO in Scenario 1, it suffered from decreased predictive accuracy and increased bias compared to both DSRIG and the LASSO in Scenario 2.

Table 3.2 records the proportion of runs for which each method (DSRIG, SRIG) was preferred, as well as the absolute median difference between the two models when the predictor graphs were estimated. DSRIG was preferred in a majority of runs and when chosen had a larger performance benefit (larger absolute median difference) than those runs for which SRIG was preferred. A greater discrepancy between the two models can be seen in Scenario 2, both in the proportion of runs preferred and in the absolute median difference. Figure 3.3 plots the difference (SRIG-DSRIG) for both the $\ell_2$-distance and RPE across all runs for the random graph scenario under Scenario 2 with sample sizes of 40/40/400. Plots with similar characteristics were obtained for all other combinations of the study parameters. Points above the horizontal zero line correspond to the models for which DSRIG outperformed SRIG; points below the line correspond to the models for which DSRIG was outperformed by SRIG. Again, it can be seen that not only does DSRIG outperform SRIG more frequently, but DSRIG also provides a greater performance benefit.
Table 3.2: Proportion of runs each method had the lowest $\ell_2$-distance or relative prediction error when the predictor graph is estimated from the training data. The median difference between SRIG and DSRIG for the $\ell_2$-distance and RPE is shown in brackets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40/40/400</td>
<td>80/80/400</td>
</tr>
<tr>
<td>DSRIG</td>
<td>0.513</td>
<td>0.347</td>
</tr>
<tr>
<td></td>
<td>(0.351)</td>
<td>(0.116)</td>
</tr>
<tr>
<td>Equivalent</td>
<td>0.302</td>
<td>0.361</td>
</tr>
<tr>
<td></td>
<td>(0.158)</td>
<td>(0.245)</td>
</tr>
<tr>
<td>SRIG</td>
<td>0.186</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>(0.158)</td>
<td>(0.212)</td>
</tr>
<tr>
<td>Scale-Free</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSRIG</td>
<td>0.513</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td>(0.673)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>Equivalent</td>
<td>0.263</td>
<td>0.376</td>
</tr>
<tr>
<td></td>
<td>(0.274)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>SRIG</td>
<td>0.224</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>(0.095)</td>
<td>(0.034)</td>
</tr>
</tbody>
</table>

3.6 Application

We demonstrate DSRIG on two real world data sets and compare performance with the LASSO and SRIG. For both scenarios all data were initially scaled such that each column of the predictor matrix ($\mathbf{X}$), and the outcome vector ($\mathbf{Y}$) had a mean of 0 and standard deviation of 1. The data were then split into 10 roughly equal segments: 8 segments on which to train the model, 1 segment to validate the model and choose the optimal value of the tuning parameter(s), and 1 segment to test the model and compare prediction accuracy across various candidate models. We fit and compared the results for all 90 possible permutations of train/validation/test sets.

As in the simulation study, the predictor graph for the SRIG and DSRIG procedures was learned on the training data using the `huge` and `huge.select` functions of the `huge` package (Zhao et al., 2015) in R (R Core Team, 2016) using the MB
Figure 3.3: Difference (SRIG-DSRIG) for 100 simulations from each of 30 random parent graphs under Scenario 2 with sample sizes of 40/40/400 for (a) $\ell_2$-distance and (b) RPE. Points above the zero line indicate when DSRIG performed better.

The STARS criterion (Liu et al., 2010) was used for the first data set. For the second data set, the rotational information criterion was used because the STARS criterion tended to result in predictor graphs with no edges. The mean square prediction error for the model chosen by the validation set applied to the independent test set was calculated as: $MSPE = \frac{1}{N_{Test}} \| \mathbf{Y} - \hat{\mathbf{Y}} \|^2_2$.

The identities of covariates with associated non-zero regression coefficient estimates ($abs(\hat{\beta}_i) > 0.01$) were also tracked.

### 3.6.1 Alzheimer’s Disease Data

Data used in the preparation of the first analysis were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). In this analysis, patients collected under all three ADNI protocols (ADNI1, ADNI2, and ADNIGO) with cross-sectional MRI were included. The objective of the analysis was two-fold: (i) to predict cognitive function from measured volumes of brain regions; and (ii) to identify which regions of the brain best contribute to this prediction. The
primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org.

Cortical reconstruction and volumetric segmentation was performed with the FreeSurfer 5.1 image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). The processing of the MRI images was conducted by Hartig et al. (2012). The technical details of these procedures are described in prior publications (Dale et al., 1999; Dale and Sereno, 1993; Fischl et al., 1999a,b; Fischl and Dale, 2000; Fischl et al., 2001, 2002, 2004a,b; Han et al., 2006; Jovicich et al., 2006; Ségonne et al., 2004). Briefly, this processing includes motion correction and averaging (Reuter et al., 2010) of multiple volumetric T1 weighted images (when more than one is available), removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Ségonne et al., 2004), automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) (Fischl et al., 2002, 2004a) intensity normalization (Sled et al., 1998), tessellation of the gray matter white matter boundary, automated topology correction (Fischl et al., 2001; Ségonne et al., 2007), and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class (Dale and Sereno, 1993; Dale et al., 1999; Fischl and Dale, 2000). The longitudinal processing capabilities of the FreeSurfer software were used in the processing of sequential scans for a single patient (Reuter et al., 2012).
The mini-mental state exam (MMSE) is a tool used by clinicians in the evaluation of cognitive functioning of psychiatric patients (Folstein et al., 1975). This exam, scored on a 30-point scale, consists of 11 questions and attempts to isolate and assess cognitive functioning from other mental functions such as mood. Higher MMSE scores correspond to higher functioning cognitive abilities. Scores greater than 27 are typically associated with normal cognitive function, whereas scores of 19-24 (mild), 10-18 (moderate), or ≤ 9 (severe), correspond to varying degrees of cognitive impairment (Kukull et al., 1994; Mungas, 1991). Yu and Liu (2016) applied the SRIG model to predict the MMSE scores of 103 subjects from the ADNI database, based on the volume measurements of 93 brain regions.

We considered 135 volume measurements (mm$^3$) obtained from the Freesurfer segmented data and restricted analysis to data available at the month 6 visit post-baseline. Of the available 696 subjects, 177 had measurements that did not pass the study’s overall quality control and a further 197 subjects did not have complete data.
We included the remaining 322 subjects in our analysis, along with their \( p = 135 \) brain region measurements and MMSE scores (range: 15–30). While these data do not represent a high-dimensional scenario where \( p > n \), many of the predictors are highly correlated and the predictor matrix \( X \) is not of full rank. Therefore, OLS would not be expected to perform well and the problem is well-suited to using regularized regression. Figure 3.4 contains the predictor graph from a single data segmentation. Note that a continuous path can be traced between all nodes in the graph, which suggests a potential violation of assumption (A2) if sparsity among the regions that predict cognitive impairment is assumed.

### 3.6.2 Blood Brain Barrier Data

In the second example, we analyzed blood brain barrier data included in the caret (Kuhn, 2016) package for R (R Core Team, 2016). More specifically, the log ratio of the concentration of a compound in the brain and the concentration of the same compound in blood for \( n = 208 \) non-proprietary literature compounds were predicted using a set of \( p = 134 \) chemical descriptors (predictors) (Mente and Lombardo, 2005). It is important to note that the predictors in this data set did not follow the multivariate Gaussian assumption. Rather, the predictors included in this analysis were binary, ordinal, and continuous predictors. However, we took a naïve approach and treated all predictors as continuous. We excluded the two predictors alert and negative as they were binary predictors with little variability (2 and 1 successes, respectively) that frequently resulted in training sets with no variability. Our analysis proceeded using the remaining \( p = 132 \) predictors.
Table 3.3: Average mean square error in prediction and mean and standard deviation for number of non-zero regression coefficients for 90 permutations of the ADNI and blood brain barrier data.

<table>
<thead>
<tr>
<th></th>
<th>LASSO</th>
<th>SRIG</th>
<th>DSRIG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADNI</strong> Mean MSPE</td>
<td>0.722</td>
<td>0.717</td>
<td>0.707</td>
</tr>
<tr>
<td>Non-zero coefficients:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean number</td>
<td>30.967</td>
<td>40.122</td>
<td>38.600</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>20.269</td>
<td>25.947</td>
<td>23.643</td>
</tr>
<tr>
<td><strong>Blood Brain Barrier</strong> Mean MSPE</td>
<td>0.538</td>
<td>0.494</td>
<td>0.491</td>
</tr>
<tr>
<td>Non-zero coefficients:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean number</td>
<td>31.689</td>
<td>47.033</td>
<td>44.444</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>12.428</td>
<td>10.299</td>
<td>10.731</td>
</tr>
</tbody>
</table>

### 3.6.3 Results of Data Analyses

Table 3.3 summarizes the model performance in terms of mean MSPE and model complexity (number of non-zero coefficients found in the final model) across all 90 data permutations for the ADNI and blood brain barrier data sets. Across both data sets the DSRIG model had the smallest mean MSPE followed by SRIG and then the LASSO. For the ADNI data, the LASSO typically chose the most sparse model with the least variability in the number of non-zero predictors, while SRIG resulted in the least sparse regression models and was the most variable in the number of non-zero coefficients selected. However, for the blood brain barrier data, while the LASSO was still the most sparse and SRIG the least sparse, we see a reversal in the variability of the number of non-zero coefficients with the LASSO having the largest variability and SRIG the smallest. For both data sets our new DSRIG model fell between the LASSO and SRIG both in terms of sparsity and in the variability of the number of non-zero coefficients.
Figure 3.5: Difference in MSPE (SRIG - DSRIG) for 90 permutations of ADNI data. Points above the zero line indicate when DSRIG performed better.

Figure 3.5 shows the differences, $SRIG - DSRIG$, between the MSPEs across each of the 90 permutations for the ADNI data. Values above the horizontal line correspond to the 27 models for which DSRIG outperformed SRIG; values below the line correspond to the 12 models for which DSRIG was outperformed by SRIG. We can see that not only did DSRIG perform better more often than SRIG, but also had a larger performance improvement (absolute mean difference 0.037 vs. 0.010).

Table 3.4 records the identity of predictors with non-zero regression coefficient estimates in at least 80 of the final models for the ADNI data. Interestingly, across the LASSO models, there was only one brain region, the right inferior lateral ventricle, that consistently was selected. There was a large overlap between the predictors commonly chosen by both DSRIG and SRIG with both models choosing three right hippocampal subfields, the right temporal pole and third ventricle. This is similar to the analysis performed by Yu and Liu (2016) which also found the right temporal pole and hippocampus to be associated with MMSE score. In our analyses,
SRIG often identified a fourth right hippocampal subfield, as well as the right hippocampus, left hippocampus and left inferior lateral ventricle in at least 80 of the 90 final models fit.

### 3.7 Discussion and Future Work

In this paper, we presented the DSRIG model which performs shrinkage and selection on components of a decomposed representation of the regression coefficients. We also derived a proximal gradient descent algorithm for parameter estimation and provided a proof of finite sample error bound. Like the SRIG model of Yu and Liu (2016), the predictor graph structure is exploited to improve the performance of regularized regression. But our model improves upon SRIG by encouraging sparsity both within and among the \( V^{(i)} \). This additional level of sparsity makes DSRIG more robust to predictor graph misspecification and to violations of assumption (A2), particularly when the predictor graph is unknown and estimated from the data.

Although results presented here do not enforce the restriction \( \max(\tau_i) \leq \xi \) (assumption (A4)), we did perform a subset of analyses with this restriction. However,
we found that prediction error was improved without it. For example, the average
MSPE after enforcing the restriction in analyzing the ADNI data was 0.710, so we
gained a 0.3% improvement through its removal. Recall that assumptions (A1)-(A6)
were required to find a finite sample error bound. (A2) is used to prove the
decomposability of $\mathcal{R}(\beta)$ while (A4) is used to bound the dual norm of $\mathcal{R}(\beta)$. When
these assumptions are not met, it does not mean DSRIG is invalid, or that a finite
error bound does not exist but rather that we are not able to derive one at this time.

DSRIG is also a highly flexible model that provides a unified framework for
fitting many other commonly used regularized regression models. When $\xi = 0$ our
DSRIG model is equivalent to SRIG. Suppose all nodes in the graph are singletons.
Then DSRIG is equivalent to the LASSO when $\tau_i = \tau \ \forall \ i = 1, \ldots, p$ and to the
adaptive-LASSO when $\xi = 0$. Furthermore, assuming an optimal decomposition,
DSRIG is equivalent to the LASSO when $\tau_i = 0 \ \forall \ i = 1, \ldots, p$. When the predictor
graph consists of a series of complete disconnected subgraphs (where each subgraph
represents a group) and $\xi = 0$ the method is equivalent to the group-LASSO. Lastly,
when the predictor graph is complete and $\xi = 0$ DSRIG is equivalent to ridge regres-
sion.

The algorithm implemented in this paper required the predictor matrix to
be represented in expanded form. For dense graphs with many edges, this procedure
can result in an augmented predictor matrix many times its original size and as such
can become computationally intensive. For the MMSE example, the predictor graph
had 135 nodes with an average of 492 edges across the 90 training sets, to result
in an augmented predictor matrix with an average of 1119 columns. Using a 2013
Mac Pro with a 6 Core 3.5GHz Intel Xeon processor running on a single core, the
average computation time for a single data split was 6.1 seconds for the LASSO, 2.1
seconds for SRIG and 52.8 seconds for DSRIG. For all three models, we considered 100 possible values for the tuning parameter $\lambda$. However, DSRIG has an additional tuning parameter $\xi$ for which we also considered 100 possible values resulting in $100 \times 100$ possible combinations of $(\lambda, \xi)$. Accordingly, part of the additional computation time was due to the additional tuning parameter rather than the expanded predictor set representation.

Future work will seek to implement a more efficient estimation algorithm to allow the method to better scale to large data sets. Regardless, DSRIG is a flexible regularization method that improves prediction, even in the presence of an imperfect predictor graph structure or violations of the Gaussian assumption.

**Acknowledgments**

This work was funded by the Natural Sciences and Engineering Research Council of Canada. Additionally, data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical
Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
Chapter 4

DSRIG: Incorporating Graphical Structure in the Regularized Modeling of SNP Data

Matthew Stephenson†, Gerarda A. Darlington†, Flavio S. Schenkel‡, E. James Squires‡, R. Ayesha Ali †

†Department of Mathematics & Statistics, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

‡Department of Animal Biosciences, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

4.1 Abstract

Genetic selection of farm animals plays an important role in genetic improvement programs. Regularized regression methods on single nucleotide polymorphism (SNP) data from a set of candidate genes can help to identify genes that are associated with the trait of interest. This complex task must also consider the relative effect sizes on the desired trait and account for the relationships among the candidate SNPs so that selection of a SNP does not promote other undesirable traits

Chapter 4 is under review for publication in Journal of Bioinformatics and Computational Biology.
through breeding. We present the Doubly Sparse Regression Incorporating Graphical structure (DSRIG), a novel regularized method for genetic selection that exploits the relationships among candidate genes to improve prediction. DSRIG was applied in the prediction of skatole and androstenone levels, two compounds known to be associated with boar taint. DSRIG was shown to provide a predictive benefit when compared to ordinary least squares (OLS) and the least absolute shrinkage and selection operator (LASSO) in a cross-validation procedure. The relative sizes of the coefficient estimates over the cross-validation procedure were compared to determine which SNPs may have the greatest impact on expression of the boar taint compounds and a consensus graph was used to infer the relationships among SNPs.

4.2 Background

Genetic selection of farm animals plays an important role in genetic improvement programs and can lead to food products that meet consumers’ nutritional needs and personal preferences. However, the identification of variants that can be used in such genetic selection can present considerable challenges when using conventional statistical models. Consider expression levels of the compounds skatole and androstenone. High expression levels of these compounds can occur in uncastrated male pigs and result in boar taint, an off odor in pork products that may render the meat inedible (Andresen, 2006). Consequently, if genetic selection can be performed to lower expression of these compounds, castration and the resulting stress endured by young male pigs may be avoided. Given a set of candidate single nucleotide polymorphisms (SNPs), interest lies in predicting skatole and androstenone expression levels and identifying SNPs that have the most potential to reduce their
expression. However, the identification of such SNPs is a complex task that must not only consider the effect size, but also ensure sufficient genetic variation exists to allow for selection in the desired direction and account for the relationships that may exist among candidate SNPs. For example, explicit selection on one SNP to reduce expression of the compounds may result in the implicit selection of a second SNP that introduces some other undesirable effect (e.g. reduced growth rate).

Suppose the set of candidate SNPs are biallelic with alleles \( A \) and \( a \) resulting in the possible genotypes \( \{AA, Aa, aa\} \). Then, assuming additive genetic effects, the SNP genotypes may be encoded as \( \{2, 1, 0\} \), respectively. Further, let the responses be derived from the classical linear model:

\[
Y = X\beta + \epsilon, \tag{4.1}
\]

where \( Y \) is an \( n \times 1 \) response vector, \( X \) is an \( n \times p \) matrix of predictors, \( \beta \) is a \( p \times 1 \) vector of regression coefficients and \( \epsilon \) is an \( n \times 1 \) vector of errors such that \( \epsilon \sim \text{Multivariate Normal}(0, \sigma^2I_n) \). When \( X \) is of full rank, the regression parameters may be estimated using ordinary least squares (OLS) such that the sum of squared residuals (SSR), i.e., the Euclidean distance between the observed and predicted response vectors, is minimized. However, this condition will be violated in the presence of multicollinearity or when \( p > n \). As such, alternative methods such as regularization are required to allow for valid parameter estimation.

Regularization can improve predictive accuracy of OLS and remains valid even when \( X \) is not of full rank (Hastie et al., 2001). A penalty term \( \mathcal{R}(\beta) \), weighted by a tuning parameter \( \lambda, \lambda \geq 0 \), is typically minimized in addition to the SSR of OLS. The penalty term, usually a norm of the regression parameter vector \( \beta \), is designed
to penalize increasingly complex models. Regularization has become an important technique in the modeling of genotype-phenotype data and has been demonstrated to result in improvement in predictive performance when compared to non-regularized counterparts (Okser et al., 2014; Ogutu and Piepho, 2014; Waldmann et al., 2013). One such commonly used regularization scheme is the least absolute shrinkage and selection operator (LASSO) where $R(\beta)$ takes the form of the $\ell_1$-norm (Tibshirani, 1996). An additional benefit of the LASSO is that it will perform variable selection by setting some of the regression parameter estimates to zero.

More recently, regularized regression models have been improved through the incorporation of structural information of the predictor set in the construction of the penalty term. Sparse Regression Incorporating Graphical structure among predictors (SRIG) assumes predictors follow a multivariate normal distribution that can be represented by an undirected graphical model and leverages this graph structure in the construction of $R(\beta)$ (Yu and Liu, 2016). SRIG operates on a decomposed representation of the regression coefficients in which each predictor contributes to the estimation of each of the $p$ regression coefficients. It encourages sparsity among the predictors that contribute to the overall estimation of $\beta$ through an $\ell_2$-groupwise penalty.

We present Doubly Sparse Regression Incorporating Graphical structure among predictors (DSRIG), a novel regularized method for genetic selection that exploits the relationships among candidate genes. DSRIG operates similarly to SRIG but adds an $\ell_1$-penalty to further encourage sparsity among the individual contributions of a predictor to each of the $p$ regression coefficients. Using cross-validation we demonstrate that DSRIG provides superior predictive accuracy to OLS and the
LASSO and that it is less variable both in the number of non-zero coefficients identified and in the magnitude of the coefficient estimates themselves. Furthermore, we show how our cross-validation procedure can be used to identify SNPs that may have the largest genetic selection impact on the phenotype of interest and we use a consensus graph to infer relationships among SNPs.

4.3 Methods

4.3.1 Doubly Sparse Regression Incorporating Graphical Structure Among Predictors

An undirected predictor graph, \( G \), may be used to model the conditional relationships among a predictor set where the nodes, \( J = \{1, \ldots, p\} \), correspond to the \( p \) random variables in \( X \). When the underlying distributions of the predictors, \( X_k, k = 1, \ldots, n \), are independent and identically distributed multivariate normal with variance-covariance matrix \( \Sigma \) and precision matrix \( \Omega = \{\omega_{ij}\}_{i,j=1,\ldots,p} = \Sigma^{-1} \), edges will exist between pairs of nodes \((i, j), i \neq j\), that are dependent conditional on all other nodes in the graph. This will occur whenever \( \omega_{ij} = \omega_{ji} \neq 0 \). In contrast, any pairs of nodes for which \( \omega_{ij} = \omega_{ji} = 0 \) are independent conditional on all other nodes in the graph and are characterized by an absence of an edge in \( G \). Nodes that share an edge are considered to be neighbours, while the neighbourhood of node \( i \in J \), \( N_i \), is defined as the union of a node and the set of all its neighbours. Further the degree of node \( i \in J \), \( d_i \), is defined to be the size of its neighbourhood, \( d_i = |N_i| \).

Define \( \Sigma_{xy} = [c_1, \ldots, c_p] \) to be the cross covariance vector between \( X \) and \( Y \). Then for observations derived from the linear model in Eq. (4.1) it can be shown
that $\beta = \Omega \Sigma_{xy}$:

$$
\beta_1 = c_1\omega_{11} + c_2\omega_{12} + \cdots + c_i\omega_{1i} + \cdots + c_p\omega_{1p},
$$

$$
\beta_2 = c_1\omega_{21} + c_2\omega_{22} + \cdots + c_i\omega_{2i} + \cdots + c_p\omega_{2p},
$$

$$
\vdots
$$

$$
\beta_p = c_1\omega_{p1} + c_2\omega_{p2} + \cdots + c_i\omega_{pi} + \cdots + c_p\omega_{pp}.
$$

(4.2)

Let $V_j^{(i)} = c_i\omega_{ji}$ for $i, j = 1, \ldots, p$ be the contribution of the $i^{th}$ predictor to the $j^{th}$ regression parameter and further define $\mathbf{V}^{(i)} = [V_j^{(i)}]_{j=1,...,p}$ to be the $p \times 1$ vector of the contributions of the $i^{th}$ predictor to each of the $p$ regression parameters. Then it follows that the regression parameter vector $\beta$ can be found by taking the sum across the $p$ contribution vectors, $\beta = \sum_{i=1}^{p} \mathbf{V}^{(i)}$. Since $\Omega$ encodes the undirected graphical structure of the predictors, each of the $\mathbf{V}^{(i)}, i = 1, \ldots, p$, will have support corresponding to the neighbourhoods $\mathcal{N}_i, i = 1, \ldots, p$, in $\mathcal{G}$.

Let $\| \cdot \|_p$ denote the $\ell_p$-norm and assume that the graph structure is known. Then the DSRIG parameter estimates can be found as the solution to:

$$
\arg \min_{\beta, \mathbf{V}^{(i)} \forall i} \left( \frac{1}{2n} \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 + \lambda \left\{ \sum_{i=1}^{p} [\tau_i \| \mathbf{V}^{(i)} \|_2 + \xi \| \mathbf{V}^{(i)} \|_1] \right\} \right), \beta = \sum_{i=1}^{p} \mathbf{V}^{(i)},
$$

(4.3)

where $\tau_i$ is a positive weight and $\xi$ is a positive tuning parameter balancing the contribution of the $\ell_1$ and $\ell_2$ components of the penalty term. The value of $\tau_i$ is calculated as $\tau_i = \sqrt{\hat{d}_i / |c_i|}$, where $\hat{d}_i$ is the estimated covariance between $\mathbf{X}_i$ and $\mathbf{Y}$, while the optimal values for $\lambda$ and $\xi$ may be tuned via cross-validation over a two-dimensional grid search.
Note that since DSRIG is not scale invariant, all variables should be normalized to have a columnwise mean of 0 and standard deviation of 1 prior to analysis. Further, DSRIG assumes a known predictor graph structure. This structure will typically need to be estimated from the data prior to carrying out the optimization. We advocate that graph estimation should be performed using only the training data so that the DSRIG estimates reflect some of the uncertainty in the predictor graph structure. Given a predictor graph structure, the support of the $V^{(i)}$’s will be known and for each combination of $(\lambda, \xi)$ the optimization problem in Eq. (4.3) can be solved using a proximal gradient descent algorithm. This procedure alternates between taking a gradient step in the negative direction of the loss function $L(\beta) = \frac{1}{2n} \| Y - X\beta \|_2^2$, and application of the proximal operator of regularizer $R(\beta) = \sum_{i=1}^{p} [\tau_i ||V^{(i)}||_2 + \xi ||V^{(i)}||_1]$.

4.3.2 Cross-Validation for Genetic Selection

Cross-validation is used in the fitting of the DSRIG model for the tuning of the paired-regularization parameters $(\lambda, \xi)$. However, cross-validation may also be useful in the identification of predictors (SNPs) most strongly associated with the outcome of interest. In cross-validation, the data are partitioned into a set of blocks of roughly equal size. Then, a portion of these blocks are used as a training set to learn both the predictor graph and parameter estimates, a portion of the blocks are used as a validation set to choose the optimal tuning parameters and, when model comparison is of interest, a portion of these blocks can be reserved as an independent test set to compare candidate models.

Following data partitioning, DSRIG and any other chosen statistical model(s)
are then fit on each of the permutations of training, validation and test sets to result in repeated estimates of each regression coefficient. The regression coefficient estimates can then be unstandardized and their empirical distributions compared to identify potential targets for genetic selection. SNPs with corresponding regression coefficient estimates that tend to be large (in absolute value), that are non-zero in a majority of the models, and do not straddle zero may be appropriate targets for genetic selection. These SNPs will be evident in boxplots of the regression coefficient estimates, where such SNPs will tend to have bounds away from zero.

When the graph structure is learned using only the training data, this cross-validation procedure will also yield several estimates of the predictor graph. However, it may be reasonable to expect edges between pairs of SNPs with strong conditional relationships to be conserved across a majority of the estimated predictor graphs. Therefore, to learn about the relationships among the SNPs, a consensus graph can be constructed by recording edges that occurred in the predictor graph in a pre-specified fraction of the estimated predictor graphs. Once a set of potential target SNPs has been identified, the consensus graph can then be used to investigate the relationships that exist among the SNPs. It is likely that SNPs in the same gene will share an edge, while edges shared between SNPs in different genes may provide useful information about gene-gene interactions. In genetic selection, when choosing between two SNPs (or two groups of SNPs) with regression coefficients of similar magnitude, it may be reasonable to choose the SNP that is more isolated in the predictor graph. SNPs that are more isolated (i.e., share fewer edges with other SNPs) may be less likely to induce unintended effects on other related biological pathways or traits.
4.3.3 Empirical Evaluation

A study involving the prediction of skatole and androstenone from SNP data was used to investigate the empirical properties of the DSRIG estimates and to determine if the incorporation of the graphical structures provides a predictive benefit over OLS and the LASSO. Animals were sourced through the provincial swine breeding associations in Quebec (CDPQ), Ontario (OSI) and Alberta (WSTA) with three breeds of pig being sampled: Duroc (n = 1631), Landrace (n = 1171), and Yorkshire (n = 1215). For each sampled pig, a fat biopsy was taken for analysis of: androstenone expression by enzyme-linked immunosorbent assay (ELISA), skatole expression by high-performance liquid chromatography (HPLC), and for extraction of genomic DNA (Huber et al., 2018). The pigs were genotyped for a set of 103 candidate SNPs using the Sequenom MassArray system at the University Health Network in Toronto. The explanatory variables of age and weight were also recorded. All analyses were performed for each breed separately.

The response variables were first log-transformed to remove skewness in the data; however, a long left tail in skatole expression levels persisted for all three breeds (see Figure 4.1). Since this study sought to identify SNPs that may result in decreased skatole production, these observations were included in the analysis. The candidate SNPs were screened according to the following:

1. Remove any SNP with more than 20% missing observations;

2. Remove any SNP with a minor allele frequency of $< 0.01$;

3. Remove any SNP with no variation.

In total, $p = 65$ SNPs were retained for Duroc, $p = 88$ SNPs for Landrace and $p = 87$ SNPs for Yorkshire. Subsequently, all observations with missing data were excluded.
resulting in final sample sizes of $n = 495$ for Duroc, $n = 660$ for Landrace and $n = 707$ for Yorkshire. The log transformed responses were then regressed on age, weight and each of their squared terms using OLS and the residuals from these models formed the response variables for all subsequent analyses. All variables were then normalized to have a columnwise mean of 0 and standard deviation of 1 prior to analysis.

The data were then partitioned into 10 blocks of roughly equal size. Eight of these blocks were reserved as a training set, one block was reserved as a validation set and one block was reserved as an independent test set. Using this structure, there were 90 permutations of these 10 blocks. For each of the 90 permutations of the data the OLS, LASSO, and DSRIG models were fit.

Predictor graphs were learned using the \texttt{huge} and \texttt{huge.select} functions of the \texttt{huge} (Zhao et al., 2015) package in \texttt{R} (R Core Team, 2016) with the Meinshausen-Buhlmann (MB) method and the STability Approach to Regularization Selection (STARS) criterion. The consensus graph was constructed by recording edges that occurred in the predictor graph for all training sets. The OLS, LASSO, and DSRIG models were fit using MATLAB version 2016b and its associated Statistics Toolbox.
Figure 4.2: Undirected consensus graph for the (a) Duroc and (b) Yorkshire breeds. Nodes, labelled V1,..., V103, represent the 103 candidate SNPs. Nodes V1, V2, and V3, corresponding to the HSD3B1 gene and its promoter, are circled. Nodes V19 and V20, corresponding to the CYB5B gene, are boxed.

(MATLAB, 2016).

To compare predictive accuracy, the final model chosen by the validation set was applied to the independent test set and the mean squared error in prediction (MSPE) was calculated as:

$$MSPE = \frac{1}{n_{Test}} \sum_{k=1}^{n} (y_k - \hat{y}_k)^2,$$  \hspace{1cm} (4.4)

where \(n_{Test}\) represents the sample size of the test set and \(y_k\) and \(\hat{y}_k\) represent the observed and predicted response for the \(k^{th}\) observation, respectively. The identity and number of SNPs with associated non-zero regression coefficient estimates (\(\text{abs}(\hat{\beta}_i) > 0.01\)) were recorded for each candidate model. To compare coefficient variability, the standard deviation of the coefficient estimates across the 90 permutations was calculated.
Table 4.1: List of 20 edges from the consensus graph for the Duroc breed. Node labels, as shown in Figure 4.2, and genes associated with the edge ends are also provided.

<table>
<thead>
<tr>
<th>Edge</th>
<th>Node</th>
<th>Gene</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V1</td>
<td>V3</td>
<td>HSD3B1-p</td>
</tr>
<tr>
<td>2</td>
<td>V2</td>
<td>V3</td>
<td>HSD3B1-p</td>
</tr>
<tr>
<td>3</td>
<td>V19</td>
<td>V20</td>
<td>CYB5B</td>
</tr>
<tr>
<td>4</td>
<td>V26</td>
<td>V28</td>
<td>CYP2A19-p</td>
</tr>
<tr>
<td>5</td>
<td>V27</td>
<td>V30</td>
<td>CYP2A19-p</td>
</tr>
<tr>
<td>6</td>
<td>V26</td>
<td>V31</td>
<td>CYP2A19-p</td>
</tr>
<tr>
<td>7</td>
<td>V28</td>
<td>V31</td>
<td>CYP2A19-p</td>
</tr>
<tr>
<td>8</td>
<td>V35</td>
<td>V36</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>9</td>
<td>V33</td>
<td>V37</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>10</td>
<td>V36</td>
<td>V38</td>
<td>CYP2E1-p</td>
</tr>
<tr>
<td>11</td>
<td>V35</td>
<td>V39</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>12</td>
<td>V35</td>
<td>V40</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>13</td>
<td>V39</td>
<td>V40</td>
<td>CYP2E1-p</td>
</tr>
<tr>
<td>14</td>
<td>V36</td>
<td>V41</td>
<td>CYP2E1-p</td>
</tr>
<tr>
<td>15</td>
<td>V38</td>
<td>V41</td>
<td>CYP2E1-p</td>
</tr>
<tr>
<td>16</td>
<td>V33</td>
<td>V42</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>17</td>
<td>V37</td>
<td>V42</td>
<td>CYP2E1-p</td>
</tr>
<tr>
<td>18</td>
<td>V35</td>
<td>V43</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>19</td>
<td>V39</td>
<td>V43</td>
<td>CYP2E1-p</td>
</tr>
<tr>
<td>20</td>
<td>V40</td>
<td>V43</td>
<td>CYP2E1-p</td>
</tr>
</tbody>
</table>

4.4 Results

Figure 4.2 depicts the consensus graphs for the Duroc and Yorkshire breeds; the nodes, labeled V1, ..., V103, correspond to the the 103 candidate SNPs genotyped. For Duroc, the consensus graph consists of 65 nodes and 47 edges; a subset of 20 of these edges are recorded in Table 4.1. The consensus graphs over the 88 nodes for Landrace and 87 nodes for Yorkshire consist of 75 and 94 edges, respectively. It was found that edges typically exist between SNPs found within a particular gene and its promoters (denoted by “p”). This pattern is consistent across all breeds.

Table 4.2 reports the average MSPE, mean, and standard deviation for the number of non-zero predictors as well as the median and maximum standard deviation of the estimated regression coefficients across the 90 permutations. Since the minimum standard deviation was 0 for all three breeds for both the LASSO and
Table 4.2: Average mean square error in prediction, mean number of non-zero regression coefficient estimates (standard deviation shown in brackets) and median standard deviation of the regression coefficients (maximum shown in brackets) for three breeds of pig and two compounds across 90 permutations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Skatole MSPE (sd)</th>
<th>Skatole Med (max)</th>
<th>Androstenone MSPE (sd)</th>
<th>Androstenone Med (max)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg Number $\hat{\beta}_i \neq 0$</td>
<td>sd($\hat{\beta}_i$)</td>
<td>Avg Number $\hat{\beta}_i \neq 0$</td>
<td>sd($\hat{\beta}_i$)</td>
</tr>
<tr>
<td><strong>Duroc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.109 46.533 (1.664) 0.095 (2.219)</td>
<td>1.663 49.133 (1.210) 0.078 (1.800)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>1.015 15.53 (17.425) 0.028 (0.188)</td>
<td>0.870 34.011 (17.749) 0.044 (0.322)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSRIG</td>
<td>1.013 16.167 (16.518) 0.018 (0.180)</td>
<td>0.854 30.700 (13.539) 0.019 (0.054)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Landrace</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>1.526 67.689 (1.929) 0.044 (2.510)</td>
<td>1.375 67.400 (2.287) 0.044 (2.363)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>0.990 11.156 (13.168) 0.014 (0.182)</td>
<td>0.971 20.067 (14.226) 0.015 (0.274)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSRIG</td>
<td>0.978 13.967 (9.058) 0.005 (0.042)</td>
<td>0.964 21.411 (12.695) 0.009 (0.052)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Yorkshire</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>1.942 71.667 (1.748) 0.077 (3.280)</td>
<td>2.569 67.867 (2.896) 0.086 (5.555)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>0.964 25.167 (21.159) 0.034 (0.331)</td>
<td>0.949 17.633 (13.631) 0.013 (0.089)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSRIG</td>
<td>0.950 25.122 (14.429) 0.008 (0.041)</td>
<td>0.941 21.267 (16.090) 0.008 (0.052)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DSRIG, the maximum equals the range for the coefficient size of these models. As expected, OLS exhibited the largest prediction error, number of non-zero coefficients as well as the largest median and maximum standard deviation of the coefficient estimates. OLS also showed the least variability in the number of non-zero regression coefficients. The LASSO offered decreased prediction error when compared to OLS and on average tended to choose models that were the most sparse among all three methods. However, the LASSO was more variable than DSRIG in the number of predictors selected and had a larger median standard deviation of the coefficient estimates. The LASSO model was also more extreme in nature than DSRIG with a much larger maximum standard deviation of the coefficient estimates. DSRIG offered the smallest prediction error and chose models with slightly more predictors than the LASSO but far fewer than OLS.

Figure 4.3 provides plots of the differences in MSPE (LASSO-DSRIG) for
Figure 4.3: Difference in MSPE (LASSO-DSRIG) across 90 permutations for all breed and compound combinations. Positive values represent permutations in which the DSRIG outperformed the LASSO with respect to MSPE.
Figure 4.4: Boxplot of regression coefficient estimates for 65 SNPs across 90 permutations for the DSRIG method. Three SNPs corresponding to the HSD3B1 gene and its promoter, are marked above boxplot by a triangle and two SNPs corresponding to the CYB5B gene are marked above boxplot by a square.

Table 4.3 reports the top seven unstandardized DSRIG regression coefficients
Table 4.3: Summary across the 90 estimates of the unstandardized DSRIG regression coefficients for the prediction of androstenone in the Duroc breed restricted to the seven SNPs that had the largest absolute median values across the 90 permutations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HSD3B1</td>
<td>0.031</td>
<td>0.159</td>
<td>0.187</td>
<td>0.180</td>
<td>0.217</td>
<td>0.278</td>
</tr>
<tr>
<td>HSD3B1</td>
<td>0.000</td>
<td>0.127</td>
<td>0.224</td>
<td>0.228</td>
<td>0.335</td>
<td>0.472</td>
</tr>
<tr>
<td>HSD3B1-p</td>
<td>0.000</td>
<td>0.090</td>
<td>0.114</td>
<td>0.111</td>
<td>0.137</td>
<td>0.201</td>
</tr>
<tr>
<td>CYB5B</td>
<td>-0.083</td>
<td>0.000</td>
<td>0.088</td>
<td>0.078</td>
<td>0.147</td>
<td>0.185</td>
</tr>
<tr>
<td>CYB5B</td>
<td>0.102</td>
<td>0.173</td>
<td>0.228</td>
<td>0.250</td>
<td>0.333</td>
<td>0.418</td>
</tr>
<tr>
<td>EPHX2</td>
<td>-0.441</td>
<td>-0.215</td>
<td>-0.171</td>
<td>-0.166</td>
<td>-0.082</td>
<td>0.000</td>
</tr>
<tr>
<td>CYB5A-p</td>
<td>0.000</td>
<td>0.144</td>
<td>0.205</td>
<td>0.177</td>
<td>0.249</td>
<td>0.388</td>
</tr>
</tbody>
</table>

with the largest absolute median values across the 90 permutations for the prediction of androstenone in Duroc pigs. Figure 4.4 presents boxplots for the regression coefficients across the 90 permutations for all 65 SNPs in the prediction of androstenone for the Duroc breed. The largest absolute median regression coefficient was associated with one of the two SNPs in the cytochrome b5 Type B (CYB5B) gene. Interestingly, the trio of SNPs corresponding to the hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1 (HSD3B1) gene and its promoter (corresponding to nodes V1, V2, and V3 in the consensus graph) all had large associated regression coefficients; these nodes correspond to the first three boxplots in Figure 4.4 and are circled in Figure 4.2.

4.5 Discussion

DSRIG is a novel regularization method that incorporates the graphical structure among the predictors to provide superior prediction results in genotype-phenotype modelling. In the empirical evaluation, DSRIG provided a predictive benefit, identified a larger number of predictors and was less variable in the number of predictors selected compared to the LASSO. Furthermore, DSRIG had a smaller
median and maximum standard deviation of the individual coefficients in the sampling procedure. These findings suggest that DSRIG has added value in the building of a genotype-phenotype predictive model and in the use of variable selection when compared to the LASSO.

DSRIG is a flexible model that allows the fitting of many other commonly used regularization schemes under the same framework. For example, DSRIG is equivalent to SRIG for $\xi = 0$. If, further, the predictor graph is complete, then DSRIG is equivalent to ridge regression. However, when all nodes in the predictor graph are singletons and $\tau_i, i = 1,...,p$, is held constant, then DSRIG is equivalent to the LASSO. When the predictor graph consists of complete disconnected subgraphs and $\xi = 0$, DSRIG is equivalent to the group-LASSO model (Yuan and Lin, 2006).

From the empirical studies performed on the boar taint example, a few potential targets to reduce androstenone expression in Duroc can be identified. One potential target could be the three SNPs associated with the HSD3B1 gene and its promoter. All three of these SNPs had large regression coefficients and, for the Duroc breed, they shared edges only among each other in the consensus graph. Another potential target could be the two SNPs located in the CYB5B gene. One of these SNPs had the largest median regression coefficient among all SNPs while the other SNP had a moderately large median coefficient (though its empirical distribution straddled zero). For the Duroc breed, these two SNPs were neighbours with each other but disconnected from other SNPs in the consensus graph. The findings of this computational analysis are supported by previous scientific studies that have found the HSD3B1 and CYB5B genes to be important in the metabolism of androstenone (Zadinová et al., 2016).

In this analysis, the size of the predictor set, and consequently the predictor
graph, for each breed of pig differed, thereby necessitating breed-specific analyses. Furthermore, the two compounds, skatole and androstenone, were each modelled separately. Future research could investigate how these methods may pool information across breeds and response variables to identify a common set of SNPs that reduce the presence of boar taint for all breeds. Lastly, although the responses here were expression levels of skatole and androstenone, boar taint itself is a binary phenotype. The authors are currently extending DSRIG to the binary response setting. Regardless, the ability to incorporate the graphical structure among predictors makes DSRIG a useful tool in the modeling of genotype-phenotype data, both for the creation of predictive models and for the identification of potential genetic targets.

Acknowledgments

The authors would like to acknowledge the Natural Sciences and Engineering Research Council for providing funding for this work as well as the Ontario Ministry of Agriculture, Food and Rural Affairs and the Agricultural Adaptation Council for providing funding associated with the creation of the marker set.
Chapter 5

DSLRIG: Leveraging Predictor Structure in Logistic Regression

Matthew Stephenson†, R. Ayesha Ali∗†, Gerarda A. Darlington†,
Flavio S. Schenkel‡, E. James Squires‡

†Department of Mathematics & Statistics, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada
‡Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

5.1 Abstract

Previous research has demonstrated that predictive performance can be improved whenever an undirected graph can be built over the set of predictors for a continuous response and the neighbourhood structure is exploited. These methods are extended to a binary outcome in the new doubly sparse logistic regression incorporating graphical structure among predictors (DSLRIG) model. DSLRIG uses a decomposed representation of the regression parameters and encourages sparsity both within and among the groups that contribute to estimation. Compared to conventional regularized methods, DSLRIG offers improved predictive performance and

Chapter 5 has been submitted for review for publication in Communications in Statistics - Simulation and Computation.
identification of the true non-zero regression coefficients, while retaining their relative ranking.

5.2 Introduction and Background

Logistic regression serves as a simple yet effective technique in the classification of a binary outcome, $Y$, from a set of input features $X$. For example, consider the problem of predicting whether the meat from a particular boar will be tainted. Boar taint is an off odor that can occur in pork products and is linked with the expression of the compound androstenone (Andresen, 2006). When androstenone levels exceed 1000 ng/g, boar taint will result and meat products will be unfavourable to consumers (Desmoulin et al., 1982). Castration can mitigate the expression of androstenone and is often used in industry to prevent this undesirable trait. However, if one could identify a set of single nucleotide polymorphisms (SNPs) associated with boar taint, then such SNPs could be incorporated into a genetic selection program instead of using castration to avoid boar taint, thereby improving animal welfare. In a regression context, the set of candidate SNPs would form the predictor set (input features) and could be linked to the response (presence of boar taint) through an unknown parameter vector, $\beta$, with interest lying in the estimation of these parameters. Given the (often) high-dimensional nature of SNP data, and the fact that the feature set often suffers from multicollinearity, the predictor matrix, $X$, may not be of full rank, which challenges conventional parameter estimation methods.

Wu et al. (2009) demonstrated that combining logistic regression with the least absolute shrinkage and selection operator (LASSO), i.e., adding an $\ell_1$ penalty to parameter estimation, is useful for modelling a binary outcome from SNP data.
More recently, Yu and Liu (2016) and Stephenson et al. (2018a) demonstrated that the structural information learned by building an undirected graph over the set of predictors can be leveraged to improve prediction of a continuous response. These latter models operate on a decomposition of the regression parameters where it can be shown that the regression parameter associated with a particular variable is comprised of a set of contributions arising from each of the variables with which it shares an edge in the predictor graph. Yu and Liu (2016) encourage sparsity among the predictors that make a contribution towards the overall estimation of $\beta$ through an $\ell_2$-group-wise penalty and Stephenson et al. (2018a) add an $\ell_1$ penalty to further encourage sparsity among the individual contributions of a particular predictor to $\beta$.

Shrinkage and selection at the group level will encourage nodes (SNPs) that share an edge in the predictor graph to have associated regression coefficients that are zero (or non-zero) together. Shrinkage and selection within a group will allow for greater flexibility. This additional level of sparsity may be important when the graph structure is unknown and must be estimated from the data. Such a mechanism may allow for the shrinkage of small contributions associated with false positive edges in the estimated predictor graph as well as allow for nodes that share an edge to have both zero and non-zero associated regression coefficients.

In this paper, we extend these ideas to the binary outcome setting in the new doubly sparse logistic regression incorporating graphical structure among predictors (DSLRIG) model. DSLRIG offers improved predictive performance when compared to the LASSO as well as an improved ability to both identify and maintain the ranking of the true non-zero regression coefficients. Section 5.3 introduces notation and presents our new DSLRIG model. Simulation study results are presented in Section 5.4, while Section 5.5 provides the results of an application to real world
data. Finally, Section 5.6 provides a general discussion and outlines areas for future research.

5.3 Notation and Methods

5.3.1 Notation

Let \( Y_k, k = 1, \ldots, n \), be a binary response for the \( k^{th} \) individual with corresponding \( 1 \times p \) covariate vector, \( X_k \). Let \( \beta \) be a \( p \times 1 \) vector of regression parameters. Then the logistic regression model may be defined as:

\[
g_\beta(X_k) = \log \left( \frac{P(Y_k = 1|X_k)}{1 - P(Y_k = 1|X_k)} \right) = X_k \beta, \tag{5.1}\]

where \( g_\beta(X_k) \) corresponds to the logit link function. Equation (5.1) can be rearranged to solve for \( P(Y_k = 1|X_k) \) as:

\[
\pi_k = P(Y_k = 1|X_k) = \frac{1}{1 + \exp(-X_k \beta)}. \tag{5.2}\]

Parameter estimates for the logistic regression model can be found by minimizing the negative log likelihood function:

\[
\ell(\beta) = -\sum_{k=1}^{n} \left[ y_k \log(\pi_k) + (1 - y_k) \log(1 - \pi_k) \right], \tag{5.3}\]

using an iteratively reweighted least squares method or gradient descent.

Let \( G \) be the undirected predictor graph over the set of nodes \( J = \{1, 2, \ldots, p\} \) corresponding to the \( p \) predictors in \( X \). Assume that the observations \( X_k, k = \)
1, \ldots, n, are independent and identically distributed multivariate normal with variance-covariance matrix $\Sigma$ and precision matrix $\Omega = \Sigma^{-1}$, comprised of elements $\omega_{ij}$. Then edges will exist between any pair of nodes $(i, j)$, $i \neq j$ for which $\omega_{ij} = \omega_{ji} \neq 0$; such nodes are dependent conditional on all other nodes in $\mathcal{G}$. When $\omega_{ij} = \omega_{ji} = 0$, $i \neq j$, nodes $i$ and $j$ are independent conditional on all other nodes in the graph, which results in the absence of the corresponding edge $(i, j)$ in $\mathcal{G}$. Define the **neighbourhood** of $i \in \mathcal{J}, \mathcal{N}_i$, to be the union of node $i$ and the set of nodes $j \in \mathcal{J}$ such that $\omega_{ij} \neq 0$. Then the **degree** of node $i$ may be defined as the size of its neighbourhood, $d_i = |\mathcal{N}_i|$.

### 5.3.2 Methods

In the case of a quantitative outcome under the classic linear model, $Y = X\beta + \epsilon$, it can be shown that the regression coefficients $\beta = \Omega \Sigma_{xy}$, where $\Sigma_{xy}$ is defined to be the cross-covariance vector between $X$ and $Y$. In practice, each vector in $X$ is typically standardized a priori to avoid scale effects. Since $\omega_{ij} \neq 0 \forall j \in \mathcal{N}_i$ it follows that node $i$ will contribute to each of the regression parameters, $\beta_j$, so long as the appropriate entries in $\Sigma_{xy}$ are non-zero. For any node $j \notin \mathcal{N}_i$, $\omega_{ij} = 0$ and consequently node $i$ will not contribute to the estimation of $\beta_j$. The sparse regression incorporating graphical structure among predictors (SRIG) model (Yu and Liu, 2016) and its doubly sparse counterpart, DSRIG (Stephenson et al., 2018a), exploit this relationship and use the neighbourhoods of the undirected predictor graph to define a set of (possibly) overlapping groups. SRIG encourages sparsity among the set of neighbourhoods that contribute to the overall estimation of $\beta$, while DSRIG additionally encourages within-neighbourhood sparsity.

The natural interpretation of the decomposition of the linear model regression coefficients, as the product of the precision matrix and the cross-covariances, does
not easily extend to the logistic regression setting; although it still may be reasonable to use the neighbourhoods of the undirected predictor graph as a set of overlapping groups. In the logistic regression setting, let the neighbourhoods, $N_i, i \in J$, represent a set of (possibly) overlapping groups. As in Obozinski et al. (2011) and Rao et al. (2013), let $\beta$ be decomposed as:

$$
\begin{align*}
\beta_1 &= V_1^{(1)} + V_1^{(2)} + \cdots + V_1^{(i)} + \cdots + V_1^{(p)}, \\
\beta_2 &= V_2^{(1)} + V_2^{(2)} + \cdots + V_2^{(i)} + \cdots + V_2^{(p)}, \\
&\vdots \\
\beta_p &= V_p^{(1)} + V_p^{(2)} + \cdots + V_p^{(i)} + \cdots + V_p^{(p)},
\end{align*}
$$

(5.4)

where $V^{(i)} = [V_1^{(i)}, V_2^{(i)}, \ldots, V_p^{(i)}]^T$ consists of the set of contributions from $N_i$ to $\beta$ with support $N_i$.

Let $\| \cdot \|_p$ represent the $\ell_p$ norm. Then the new DSLRIG model can be proposed with parameter estimates found as the solution to:

$$
\arg \min_{\beta, V^{(i)} \forall i} \left( \ell(\beta) + \lambda \left\{ \sum_{i=1}^{p} [\tau_i \|V^{(i)}\|_2 + \xi \|V^{(i)}\|_1] \right\} \right), \quad \beta = \sum_{i=1}^{p} V^{(i)},
$$

(5.5)

for tuning parameters $\lambda \geq 0$ and $\xi \geq 0$ and group-specific weight $\tau_i = \sqrt{d_i}$. The tuning parameter $\xi$ will balance the contributions of the $\ell_1$ and $\ell_2$ components of the penalty function. The $\ell_1$ component of Equation (5.5) will encourage sparsity among the contributions shrinking small individual $V_j^{(i)}$ to zero; while the $\ell_2$ component of Equation (5.5) will encourage sparsity among the neighbourhoods that contribute to $\beta$ by shrinking entire vectors $V^{(i)}$ to zero. Just as DSRIG is equivalent to SRIG (Yu and Liu, 2016) when $\xi = 0$, DSLRIG reduces to a sparse logistic regression incorporating
graphical structure (SLRIG) when $\xi = 0$. If all nodes $i \in J$ are singletons; i.e., there are no edges in $G$, then the optimization problem will reduce to the standard logistic LASSO.

Parameter Estimation

Let $\tilde{V}$ be a concatenation of all non-zero components of $V^{(i)}$, $i = 1, \ldots, p$ and $\tilde{X}$ be an augmented predictor matrix consisting of replicated columns of the original predictor matrix $X^{(j)}$, once for each neighbourhood in which $j$ belongs. Consequently, $\tilde{V}$ will be a $\sum_{i=1}^{p} d_i \times 1$ vector, while $\tilde{X}$ will be an $n \times \sum_{i=1}^{p} d_i$ matrix and the linear predictor $X\beta$ may be rewritten in an expanded form as $\tilde{X}\tilde{V}$. Then the DSLRIG optimization problem in Equation (5.5) may be written as:

$$\arg\min_{\beta, V^{(i)}_{\forall i}} \left( \ell(\beta) + \lambda \left\{ \sum_{i=1}^{p} \left[ \tau_i \|\tilde{V}^{(i)}\|_2 + \xi \|\tilde{V}\|_1 \right] \right\} \right), \beta = \sum_{i=1}^{p} V^{(i)},$$

where $\tilde{V}^{(i)}$ consists of only the non-zero components of $V^{(i)}$.

The optimization problem in Equation (5.6) can be separated into two components: (i) $\ell(\beta)$ (see Equation (5.3)), a smooth function assessing the model fit to the data; and (ii) $R(\beta) = \sum_{i=1}^{p} \left[ \tau_i \|\tilde{V}^{(i)}\|_2 + \xi \|\tilde{V}\|_1 \right]$, a non-smooth penalty function penalizing for increasingly complex models. Following Rao et al. (2014), a proximal gradient descent method that iterates between taking a gradient step in the negative direction of $\ell(\beta)$ and application of the proximal operator of $R(\beta)$ can be used to solve Equation (5.6). Let $\left( \tilde{V}^{(i)} \right)^{(r)}$ be the parameter estimates for iteration $r$ and let $\tilde{V}_{\forall}$ be the intermediary result following the gradient step to which the proximal operator will be applied. Then the proximal operator consists of first a soft thresholding
of the $j = 1, \ldots, \sum_{i=1}^{p} d_i$ individual elements of $\tilde{\nabla}$:

$$\tilde{V}_j^* = \begin{cases} 
\text{sign}(\tilde{V}_j) - \lambda \xi & \text{for } |\tilde{V}_j| > \lambda \xi \\
0 & \text{otherwise}
\end{cases}, \quad (5.7)$$

followed by a group soft thresholding across each of the $i = 1, \ldots, p$ neighbourhood contribution vectors, $\tilde{V}^{(i)}$ to $\beta$:

$$\left(\tilde{V}^{(i)}\right)^{r+1} = \begin{cases} 
\tilde{V}^{(i)}^* \left(\|\tilde{V}^{(i)}^*\|_2 - \lambda \tau_i\right) & \text{for } \|\tilde{V}^{(i)}^*\|_2 > \lambda \tau_i \\
0 & \text{otherwise}
\end{cases}. \quad (5.8)$$

From these steps it can be seen how DSLRIG first encourages sparsity within a neighbourhood’s contributions through the element-wise soft thresholding and then encourages sparsity among the neighbourhoods through the group soft thresholding.

### 5.3.3 Boar Taint Data

Data were collected through the provincial swine breeding associations in Quebec (CDPQ), Ontario (OSI) and Alberta (WSTA) on three breeds of pigs: Duroc ($n = 1631$), Landrace ($n = 1171$), and Yorkshire ($n = 1215$). A fat biopsy on each pig was used for analysis of androstenone expression by enzyme-linked immunosorbent assay (ELISA) and extraction of genomic DNA (Huber et al., 2018). Genotypes were recorded using the Sequenom MassArray system at the University Health Network in Toronto for a set of 103 biallelic SNPs with possible genotypes of $\{AA, Aa, aa\}$ that were encoded as $\{2, 1, 0\}$. Age and body weight were also recorded for each pig.

The SNPs were screened by removing any SNP that had more than 20%
missing observations, a minor allele frequency of less than 0.01, or no variation. After screening, there remained 65 SNPs for Duroc, 88 SNPs for Landrace, and 87 SNPs for Yorkshire. Any observation that was incomplete was excluded resulting in sample sizes of 513 for Duroc, 670 for Landrace, and 716 for Yorkshire.

5.4 Simulation Study

The simulation study restricts analysis to only the Duroc breed and makes use of parameter estimates from a previous analysis that treated androstenone as a continuous response and included the compound skatole (Stephenson et al., 2018b). There were only 495 observations with measurements on androstenone, skatole and the 65 SNPs, so the simulation study uses a sample size of 495.

5.4.1 Data Generation

The predictor sets were generated from a multivariate normal distribution using the `rmvnorm` package (Genz et al., 2017; Genz and Bretz, 2009) in R (R Core Team, 2016) with mean vector $\mu$ and variance-covariance matrix $\Sigma$ taken from the standardized empirical distribution of the 495 complete observations for the Duroc breed. The regression parameters associated with the $p = 65$ SNPs were first initialized as the median estimate from a set of 90 cross-validated DSRIG parameter estimates of Stephenson et al. (2018b). To enforce sparsity among the true regression parameters, simulation regression parameters were set to zero if the first or third quartile of the cross-validated estimates were zero. This resulted in the true regression parameter vector having 24 non-zero and 41 zero entries. Finally, the simulation regression parameter vector was multiplied by 3 to allow for better separation of the
estimated probability of an outcome of interest from a logistic regression model.

The outcome was first generated as a continuous response from the linear model $Y^*_k = X_k \beta + \epsilon_k, k = 1, \ldots, n$ where $\epsilon_k \sim \text{Normal}(0, \sqrt{0.85035})$ ($\sigma$ estimated from the boar taint data using an ordinary linear regression). Subsequently, the outcome was dichotomized as:

$$Y_k = \begin{cases} 0, & \text{for } Y^*_k \leq 0 \\
1, & \text{for } Y^*_k > 0 \end{cases}.$$

The threshold of 0 was chosen as it (roughly) corresponded to an androstenone level of 1000 ng/g for the unstandardized data and resulted in the same splitting that was seen in the real data with approximately 50% of observations exceeding this threshold.

The simulated data were split into 3 sets: a training set that was used to learn the regression parameters, a validation set that was used to choose the optimal value of the tuning parameter(s) and a test set that was used to compare the various candidate models. The training and validation sets were chosen to be of equal size, with ten different sample sizes being considered ranging from $n = 100$ to $n = 1,000$ in increments of 100. The sample size of the test set was held constant at $n = 10,000$ to avoid any small sample effects. A total of 100 data sets were generated for each sample size.

### 5.4.2 Model Fitting

The predictor graphs taken as input for the SLRIG and DSLRIG models were learned using the `huge` (Zhao et al., 2015) package in R (R Core Team, 2016) based on the Meinhausen-Bühlman (MB) method (Meinshausen and Bühlmann, 2006)
and with the best graph being selected using the stability approach to regularization selection (STARS) criterion (Liu et al., 2010). The following models were fit: logistic regression, logistic regression with a LASSO penalty, SLRIG and DSLRIG. The logistic regression and LASSO models were fit using MATLAB version 2016b and its associated Statistics Toolbox (MATLAB, 2016), while the proximal gradient methods for the SLRIG and DSLRIG models were implemented in MATLAB using code adapted from Cox (2014).

For the regularized models, 100 values were considered for the tuning parameter $\lambda$ with the values for the LASSO model being chosen using MATLAB’s default method. For the SLRIG and DSLRIG models the maximum value for $\lambda$, $\lambda_{\text{max}}$, was specified as the minimum value of $\lambda$ that resulted in an estimated coefficient vector of $\hat{\beta} = 0$ (for $\xi = 0$). The remaining values for $\lambda$ were equally spaced on a logarithmic scale with $\lambda_{\text{min}} = 0.01 \lambda_{\text{max}}$. For the DSLRIG tuning parameter $\xi$, 100 equally spaced values on a linear scale were considered between 0 and $\xi_{\text{max}}$, where $\xi_{\text{max}}$ was set to be just large enough to never be chosen. The greatest classification accuracy in the validation set was used to choose the best tuning parameter(s). Since the number of observations in the validation set was fixed, it was possible for different tuning parameters to result in the same classification accuracy. Therefore, two scenarios were considered for tie-handling: (i) choosing the model with $\min \| \hat{\beta} \|_2$ (most regularized); and (ii) choosing the model with $\max \| \hat{\beta} \|_2$ (least regularized).

For each individual run of the simulation study the following statistics were recorded. The proportion of observations misclassified for the final chosen model applied to the test set was used to compare predictive accuracy. The proportion of regression coefficients correctly classified as non-zero and the proportion of regression coefficients correctly classified overall (as zero or non-zero) were used to compare
model selection accuracy. Lastly, the mean absolute rank differences of the coefficients were used to compare a model’s ability to correctly rank the coefficients. Mean absolute rank difference was calculated by first ranking the magnitudes of the true and estimated regression coefficients, respectively, in order of decreasing value. Second, for each coefficient, the difference between its true and estimated ranks was calculated. The mean absolute rank difference was computed both over the set of the true non-zero regression coefficients and over all the coefficients. Note that although the outcome was generated on a continuous scale and subsequently dichotomized, previous research has shown that the regression coefficients of the logistic regression model will simply be a rescaling of the linear regression coefficients (Moser and Coombs, 2004). As such the identities of the non-zero regression coefficients and their ranks will remain unchanged. Finally, the median value for each statistic across all runs for a given sample size was recorded and used to compare the candidate models.

5.4.3 Results

Table 5.1 records the median proportion of observations misclassified across the 100 iterations of the simulation study by sample size. Regardless of which method was used for tie-handling, the misclassification rates for the regularized models were smaller than that of a conventional logistic regression model, particularly when the sample size was small. Furthermore, the two models that incorporate the graphical structure of the predictors, SLRIG and DSLRIG, demonstrated improved prediction over the LASSO model.

Figure 5.1(a) plots the median proportion of the true non-zero regression coefficients correctly estimated to be non-zero, while Figure 5.1(b) plots the median proportion of correctly estimated coefficients (zero or non-zero), i.e., the overall match
Table 5.1: Median proportion of observations misclassified in the test set across 100 iterations for ten different sample sizes of training and validation sets.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Logistic Regression</th>
<th>min $|\hat{\beta}|_2$</th>
<th>max $|\hat{\beta}|_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LASSO SLRIG DSLRIG</td>
<td>LASSO SLRIG DSLRIG</td>
<td>LASSO SLRIG DSLRIG</td>
</tr>
<tr>
<td>100</td>
<td>0.373 0.303 0.298</td>
<td>0.298 0.298 0.297</td>
<td>0.304 0.298 0.297</td>
</tr>
<tr>
<td>200</td>
<td>0.314 0.272 0.268</td>
<td>0.270 0.257 0.256</td>
<td>0.272 0.268 0.269</td>
</tr>
<tr>
<td>300</td>
<td>0.282 0.259 0.256</td>
<td>0.247 0.244 0.243</td>
<td>0.251 0.248 0.247</td>
</tr>
<tr>
<td>400</td>
<td>0.266 0.250 0.247</td>
<td>0.248 0.244 0.243</td>
<td>0.251 0.248 0.247</td>
</tr>
<tr>
<td>500</td>
<td>0.258 0.245 0.243</td>
<td>0.244 0.241 0.239</td>
<td>0.245 0.242 0.243</td>
</tr>
<tr>
<td>600</td>
<td>0.252 0.241 0.239</td>
<td>0.240 0.239 0.239</td>
<td>0.241 0.239 0.239</td>
</tr>
<tr>
<td>700</td>
<td>0.249 0.239 0.238</td>
<td>0.239 0.239 0.239</td>
<td>0.239 0.239 0.239</td>
</tr>
<tr>
<td>800</td>
<td>0.246 0.237 0.237</td>
<td>0.237 0.236 0.236</td>
<td>0.237 0.236 0.236</td>
</tr>
<tr>
<td>900</td>
<td>0.243 0.235 0.236</td>
<td>0.236 0.235 0.235</td>
<td>0.235 0.234 0.235</td>
</tr>
<tr>
<td>1000</td>
<td>0.242 0.235 0.235</td>
<td>0.235 0.235 0.235</td>
<td>0.235 0.234 0.234</td>
</tr>
</tbody>
</table>

Figure 5.1: (a) Median proportion of the true non-zero regression coefficients correctly estimated to be non-zero; (b) Overall match rate, given by the median proportion of correctly classified coefficients (zero or non-zero). Proportions based on 100 iterations for ten different sample sizes.
rate. While SLRIG identified the largest proportion of the true non-zero regression coefficients, its overall match rate was the poorest among all methods. The LASSO model achieved a greater overall match rate than SLRIG, however it failed to identify a large proportion of the true non-zero regression coefficients. DSLRIG outperformed the LASSO in overall match rate while nearly matching SLRIG in the proportion of true non-zero regression coefficients identified. There is little difference in the proportion of true non-zero coefficients correctly identified between the min $\|\hat{\beta}\|_2$ and the max $\|\hat{\beta}\|_2$ for a given model; though, the most regularized model (min $\|\hat{\beta}\|_2$) tended to result in superior performance for the overall match rate.

Figure 5.2(a) plots the median mean absolute rank difference of the absolute value of the coefficients computed over the set of true non-zero regression coefficients, while Figure 5.2(b) plots this difference computed over all coefficients. It can be seen that both SLRIG and DSLRIG tend to outperform the LASSO achieving a smaller mean rank difference across a majority of the sample sizes both over the set of non-zero coefficients as well as over the set of all coefficients. DSLRIG tended to have a smaller mean rank difference compared to the SLRIG models when considering all coefficients but there was little difference between the two models when considering only the non-zero coefficients.

5.5 Application

The logistic regression, LASSO, SLRIG and DSLRIG models were applied to the real world boar taint data. The explanatory variables of age and body weight were also included as potential predictors and were included in the predictor graph. The response variable, corresponding to androstenone level, was dichotomized using
Figure 5.2: Median mean absolute rank difference of the absolute value of the coefficients over the true non-zero regression coefficients in (a); and all coefficients in (b).

The 1000 ng/g threshold with 49% of Duroc, 28% of Landrace, and 25% of Yorkshire pigs with androstenone level exceeding the threshold and being deemed as tainted. All covariates were standardized to have mean 0 and standard deviation 1. Analyses were performed on each breed separately.

A cross-validation procedure was used to compare the performance of the candidate models used in the simulation study. For each breed, the data were partitioned into a single independent test observation, a validation set consisting of approximately 25% of the observations, and a training set consisting of all remaining observations. Specifically, the validation set sizes were: 125 for Duroc, 150 for Landrace, and 175 for Yorkshire. The data splitting was then repeated such that each observation had a chance to be the test observation. As in the simulation study, the predictor graphs were learned using the huge (Zhao et al., 2015) package in R (R Core Team, 2016) with the MB method (Meinshausen and Bühlmann, 2006) and the
Figure 5.3: Undirected predictor graph for the Duroc breed from one of the data splits.

The proportion of correctly classified observations for the final chosen model from the validation set applied to the independent test observation was used to compare the candidate models. The identity and number of non-zero estimated regression coefficients ($|\hat{\beta}| > 0.01$) for each of the final chosen models were recorded. The cross-validation procedure yielded multiple estimates for each of the regression coefficients. The standard deviation was then computed across the cross-validated estimates for each regression coefficient individually and the median of these standard deviations was recorded.

Table 5.2 records the boar taint classification accuracy, median standard deviation of the regression coefficients and median and interquartile range (IQR) for
Table 5.2: Comparison of classification accuracy (tainted: yes/no), median standard deviation of the cross-validated regression coefficients and median and interquartile range (IQR) for number of non-zero coefficient estimates for three breeds of pigs.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Classification</th>
<th>Logistic</th>
<th>LASSO</th>
<th>SLRIG</th>
<th>DSLRIG</th>
<th>LASSO</th>
<th>SLRIG</th>
<th>DSLRIG</th>
<th>min $|\hat{\beta}|_2$</th>
<th>max $|\hat{\beta}|_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Duroc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.655</td>
<td>0.678</td>
<td>0.690</td>
<td>0.690</td>
<td>0.680</td>
<td>0.698</td>
<td>0.702</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.16×10^{12}</td>
<td>0.111</td>
<td>0.046</td>
<td>0.041</td>
<td>0.133</td>
<td>0.057</td>
<td>0.053</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td>52</td>
<td>31</td>
<td>42</td>
<td>36</td>
<td>37</td>
<td>47</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IQR</td>
<td>2</td>
<td>15</td>
<td>17</td>
<td>19</td>
<td>22</td>
<td>14</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td><strong>Landrace</strong></td>
<td></td>
<td>0.718</td>
<td>0.725</td>
<td>0.716</td>
<td>0.718</td>
<td>0.721</td>
<td>0.727</td>
<td>0.734</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.68×10^{13}</td>
<td>0.073</td>
<td>0.039</td>
<td>0.039</td>
<td>0.090</td>
<td>0.050</td>
<td>0.049</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td>74</td>
<td>36</td>
<td>55</td>
<td>50</td>
<td>44</td>
<td>64</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IQR</td>
<td>2</td>
<td>27</td>
<td>31</td>
<td>33</td>
<td>24</td>
<td>22</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td><strong>Yorkshire</strong></td>
<td></td>
<td>0.715</td>
<td>0.753</td>
<td>0.757</td>
<td>0.757</td>
<td>0.749</td>
<td>0.753</td>
<td>0.751</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.80×10^{13}</td>
<td>0.063</td>
<td>0.020</td>
<td>0.020</td>
<td>0.086</td>
<td>0.031</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td>73</td>
<td>13</td>
<td>22</td>
<td>22</td>
<td>20</td>
<td>35</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IQR</td>
<td>2</td>
<td>15</td>
<td>22</td>
<td>21</td>
<td>17</td>
<td>26</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>
the number of non-zero coefficient estimates across the three breeds of pigs. It can be seen that the regularized models typically resulted in an improved classification accuracy when compared to the conventional logistic regression model. For both the Duroc and Landrace breeds, the DSLRIG (max $\|\hat{\beta}\|_2$) resulted in the greatest classification accuracy across all models, while the SLRIG (min $\|\hat{\beta}\|_2$) and DSLRIG (min $\|\hat{\beta}\|_2$) were tied for the greatest classification accuracy in the Yorkshire breed. The logistic regression model tended to identify a large number of non-zero coefficient estimates with little variability in the number identified but with variability across the cross-validated regression coefficient estimates tending towards infinity. The LASSO tended to result in the most sparse models with the least variability in the number of non-zero coefficients identified but had large variability in the cross-validated coefficient estimates. DSLRIG tended to result in a model that was more sparse than SLRIG but with greater variability in the number of non-zero coefficients identified. DSLRIG also tended to have the smallest variability in the cross-validated coefficient estimates.

Table 5.3 presents a summary of the DSLRIG (max $\|\hat{\beta}\|_2$) regression coefficients across the 513 estimates for the Duroc breed restricted to the five SNPs that had the largest absolute median values. SNPs found in the HSD3B1, CYB5A-promoter (2 SNPs), CYB5B, and BAT1 genes were found to have large associated regression coefficients. The regression coefficient estimates associated with these SNPs were found to be non-zero in 87.1–100% of the data splits. Body weight was also found to be an important predictor of boar taint with a non-zero estimate in 99.8% of the data splits and a median regression coefficient estimate of 1.343.
Table 5.3: Summary across the 513 estimates of the DSLRIG (max $||\hat{\beta}||_2$) regression coefficients for the Duroc breed restricted to the five covariates that had the largest absolute median values.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Min.</th>
<th>1st.Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd.Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSD3B1</td>
<td>0.000</td>
<td>0.236</td>
<td>0.289</td>
<td>0.289</td>
<td>0.341</td>
<td>0.544</td>
</tr>
<tr>
<td>CYB5A-promoter(1)</td>
<td>0.000</td>
<td>0.154</td>
<td>0.251</td>
<td>0.237</td>
<td>0.327</td>
<td>0.568</td>
</tr>
<tr>
<td>CYB5B</td>
<td>0.010</td>
<td>0.175</td>
<td>0.232</td>
<td>0.252</td>
<td>0.323</td>
<td>0.673</td>
</tr>
<tr>
<td>BAT1</td>
<td>0.000</td>
<td>0.107</td>
<td>0.213</td>
<td>0.210</td>
<td>0.318</td>
<td>0.557</td>
</tr>
<tr>
<td>CYB5A-promoter(2)</td>
<td>-0.555</td>
<td>-0.294</td>
<td>-0.205</td>
<td>-0.202</td>
<td>-0.097</td>
<td>0.000</td>
</tr>
</tbody>
</table>

5.6 Discussion

In this paper, the SRIG and DSRIG methods are extended to a binary response setting. These models leverage information learned by building an undirected graph over the predictor set to improve prediction accuracy and model selection performance in the modelling of a binary outcome. The neighbourhoods of the undirected graph are used to form a set of overlapping groups and sparsity is encouraged both within and among the groups that contribute to the overall estimation of $\beta$. Through simulation it was found that both SLRIG and DSLRIG resulted in improved predictive performance as well as an improved ability to both identify and maintain the ranking of the true non-zero regression coefficients when compared to the LASSO.

From the simulation study, using max $||\hat{\beta}||_2$ for tied models resulted in a greater proportion of true non-zero regression coefficients being correctly estimated as non-zero and a smaller mean rank difference when compared to that using min $||\hat{\beta}||_2$. However, the overall match rate tended to be better when using min $||\hat{\beta}||_2$ for tied models. Using max $||\hat{\beta}||_2$ is less restrictive allowing more opportunity for true active coefficients to be picked up by the model, but does so at the expense of also being more likely to include false positives (true non-zeros that are deemed active). In contrast, using min $||\hat{\beta}||_2$ is more regularized, allowing for more coefficients to be zeroed out, but
is also more likely to include false negatives. Consequently, in this sparse regression setting where the majority of coefficients are likely to equal zero, it is not surprising that overall match rate was highest using \( \min \| \hat{\beta} \|_2 \). Ultimately, the choice of which method to use will depend on the modelling objective.

Note that in the simulation study the regression parameters were multiplied by 3 because when left at their original values the estimated probabilities of the outcome tended to fall very close to 0.5 for a majority of the observations. These estimated probabilities led to poor predictive performance for all models. Multiplying the coefficients by 3 led to better separation of the estimated probabilities of the outcome, thereby resulting in an improved predictive performance in the simulation study. However, the results and general trends when the coefficients were left at their original values are similar to those found in the simulation study (see Appendix B for additional results).

In the application to the real data, it was found that SNPs in the HSD3B1, CYB5A-promoter (2 SNPs), CYB5B, and BAT1 genes had large associated regression coefficients. These findings mimic the results of the analogous study that treated androstenone as a continuous variable and SNPs in the HSD3B1 and CYB5B genes were found to have large associated regression coefficients (Stephenson et al., 2018b). Importantly, previous scientific studies have also demonstrated that these genes are key to the metabolism of androstenone (Zadinová et al., 2016). Therefore, the HSD3B1 and CYB5B genes may serve as suitable targets for genetic selection to prevent boar taint without the need for castration.

The algorithm presented in Section 5.3.2 relies on replication of the predictors once for each neighbourhood in which they belong. If the graph is dense with many edges, this replication can result in an augmented predictor matrix that is many
times its original size and parameter estimation that becomes computationally intensive. Future work will seek to develop estimation algorithms that are more efficient to allow for better scaling to large data sets with dense predictor graphs.

Acknowledgments

The authors would like to extend our thanks to the Natural Sciences and Engineering Research Council for providing funding to conduct this research. We also thank the Ontario Ministry of Agriculture, Food and Rural Affairs and the Agricultural Adaptation Council for providing funding used for data collection in the creation of the marker set.
Chapter 6

Discussion, Conclusions and Future Work

In this thesis, a novel regularization scheme that incorporates the graphical structure among predictors was proposed and evaluated. In general, this regularization scheme works by building an undirected graph over the predictor set and using the resulting neighbourhoods of the graph to form a set of overlapping groups. Sparsity is then encouraged both within and among the groups that contribute to the overall estimation of the regression parameters. Ultimately, the incorporation of the graphical structure of the predictors in this way was found to improve outcome prediction and parameter estimation for both quantitative and binary outcomes under differing predictor structures.

In Chapter 3, this new regularization scheme was proposed as the DSRIG model for use in the case of a quantitative outcome and for predictors distributed as MVN. An algorithm for parameter estimation was presented and the theoretical properties were explored through the derivation of a finite sample error bound. Empirical properties of DSRIG were investigated through a comprehensive simulation study and application to two real world data sets.

It was demonstrated that DSRIG has improved predictive performance and decreased bias in parameter estimation compared to OLS, the LASSO, and SRIG. DSRIG was also found to be more robust than SRIG to violations of the assumption that a predictor associated with a non-zero regression coefficient is only a neighbour
to other predictors also associated with non-zero regression coefficients Assumption (A2). Additionally, since the performance benefit of DSRIG (over SRIG) was of greater magnitude when the predictor graph was treated as unknown and estimated from the training data, DSRIG may also be more robust to errors in estimation of the input predictor graph.

In Chapter 4, DSRIG was extended beyond the case of MVN predictors and was used in the modelling of genotype-phenotype data. Here expression levels of two compounds associated with boar taint formed the outcome, while the predictors consisted of a set of biallelic SNP genotypes encoded as \{0, 1, 2\}. A cross-validation study using this real world data set was used to compare model performance. Practical guidelines for the implementation of DSRIG in the identification of target genes for genetic selection were also developed.

In the context of genotype-phenotype modelling, DSRIG resulted in improved predictive performance compared to OLS and the LASSO. Compared to the LASSO, DSRIG identified a larger number of predictors and was less variable in the number of predictors selected. Furthermore, DSRIG had a smaller median and maximum standard deviation of the individual coefficients in the sampling procedure. These findings suggest that DSRIG has added value over the LASSO for building a genotype-phenotype predictive model and for variable selection.

Finally, in Chapter 5 the SLRIG and DSLRIG models were proposed as extensions of SRIG and DSRIG to the binary outcome setting where a logistic regression model is appropriate. An algorithm for parameter estimation was developed and empirical properties of these new models were investigated through a simulation study. In addition to the simulation study, the real world data analyzed in Chapter 4 were transformed to include a binary response and model performance was assessed
via cross-validation.

Through simulation it was found that both SLRIG and DSLRIG resulted in improved predictive performance as well as an improved ability to both identify and maintain the ranking of the true non-zero regression coefficients compared to the LASSO. Similar to the case of a quantitative outcome, DSLRIG identified a larger number of predictors when compared to the LASSO and had smaller variability in the individual coefficients over the cross-validation procedure.

6.1 Analysis of Boar Taint Data

The application sections of Chapters 4 and 5 both made use of the same boar taint data set. In Chapter 4, the expressions of the two compounds, androstenone and skatole, were treated as quantitative variables. In Chapter 5, the expression levels of these compounds were dichotomized using reference values that correspond to the presence of boar taint in meat. Since less than 2% of pigs exceeded the undesirable threshold for skatole, analysis was restricted to androstenone in Chapter 5. The analyses of the boar taint data in Chapters 4 and 5 also differed with respect to how the predictors age and weight were handled. When the compounds were treated as quantitative outcomes, expression levels were first regressed on age, age$^2$, weight and weight$^2$ and the residuals were then regressed on the SNP predictors. Since this approach is not possible for a binary response, age and weight were instead included in the predictor graph along with the SNPs and were incorporated into the regularization scheme. The inclusion of the quadratic terms into the binary analysis was considered, but these terms were found to have little effect on outcome prediction and were excluded from the final analysis.
In these analyses, high expression levels of androstenone were far more likely to occur in the Duroc breed than for either Landrace or Yorkshire. It was found that 49% of Duroc, 28% of Landrace and 24% of Yorkshire pigs exceeded the 1000 ng/g threshold for androstenone expression. This is consistent with previous research that has also shown that the Duroc breed tends to have higher expression levels of androstenone compared to Landrace and Yorkshire (Xue et al., 1996). The high androstenone levels in Duroc may partially be explained by the fact that Duroc is most commonly a terminal breed and selected for traits associated with meat production such as growth, feed conversion, and meat quality (Rothschild and Ruvinsky, 2011; Sullivan, 2007). In contrast, the Landrace and Yorkshire breeds tend to be used as maternal breeds and selected for maternal traits such as litter size (Rothschild and Ruvinsky, 2011; Sullivan, 2007). As boar taint tends to be more prevalent in Duroc pigs, the identification of targets for genetic selection through investigation into SNPs associated with large regression coefficient estimates focused only on the Duroc breed.

As expected, the SNPs found to be associated with large regression coefficient estimates tended to be similar for both the linear and logistic regression analyses. A further comparison of the distributions of the cross-validated regression parameter estimates for the DSRIG model in the quantitative outcome setting and that for the DSLRIG-max∥\hat{β}∥_2 model in the binary outcome setting across all 65 SNPs can be found in Figure C.1 in Appendix C. Regardless of how the outcome was treated, there was similarity among the distributions of the estimated regression coefficients. The SNPs associated with large median regression coefficients and boxplots bounded away from zero tended to be similar for both analyses. Among the SNPs with the top 5 largest median regression coefficients (in absolute value), both analyses contained SNPs within the HSD3B1 and CYB5B genes. Importantly, the findings of
these computational analyses are well supported by previous scientific studies that have also found the HSD3B1 and CYB5B genes to be important in the metabolism of androstenone (Zadinová et al., 2016).

6.2 Future Work

There are many potential avenues for future work stemming from the contributions presented in this thesis. The fitting algorithm implemented in the estimation of both the DSRIG and DSLRIG models relies on predictor duplication to create an augmented predictor matrix that consists of replicates of the columns of the original predictor matrix once for each neighbourhood in which a given predictor belongs. The predictor duplication method results in the estimation of the individual contributions to the regression parameters, i.e., the $V_{ij}^{(i)}$, $i,j = 1,\ldots,p$. At the cost of computational efficiency, this allows for easy application of the proximal operator of these models whereby the individual elements, $V_{ij}^{(i)}$, are first soft thresholded and then followed by a subsequent group soft thresholding on the vectors of contributions, i.e., the $V^{(i)}$, $i = 1,\ldots,p$. However, it would be of benefit if an algorithm could be derived that can apply this hierarchical proximal operator directly to $\beta$ without the need for predictor duplication. Such an algorithm would be particularly useful in reducing the computational cost of these doubly sparse methods as the number of predictors increases and/or the predictor graph becomes more dense.

Secondly, the theoretical properties derived in Chapter 3 were for the particular case of a predictor set that is distributed as MVN. It is of interest to develop theoretical results that apply as the predictors diverge from a MVN distribution.
Such results would be more applicable to scenarios where the predictor set is comprised of a set of SNPs, similar to the data analyzed in Chapter 4. Additionally, these theoretical properties were only for the case of a quantitative outcome where the error component was assumed to be distributed as $\text{MVN}(0, \sigma^2 I_n)$. This allowed for the use of the chi-square distribution to find an error bound that holds with high probability. Since logistic regression does not have a similar normally distributed error component, these results do not readily extend to binary outcomes. Future work could entail deriving a finite sample error bound that holds for the binary outcome scenario through the use of different distributional assumptions.

The finite sample error bound derived in Chapter 3 was for the error in estimation. It would also be of interest to derive results for model selection consistency for both the quantitative and the binary response cases. This would provide a theoretical basis to further support the findings in Chapter 5, where it was shown that DSLRIG resulted in improved identification of the true non-zero regression coefficients when compared to the LASSO.

Lastly, in Chapters 4 and 5, the analyses were separated by breed. However, an analysis that can pool information across all breeds may be beneficial. In Chapter 5, the outcome only considered whether or not androstenone exceeded a threshold associated with the presence of boar taint, but boar taint could also result from high expression levels of skatole. Extending DSRIG/DSLRIG to the multi-task, or multivariate, setting would allow for both compounds, androstenone and skatole, to be modelled simultaneously. Finally, from a scientific perspective, experiments could be conducted to determine the effects of a genetic selection program that targets the SNPs in the HSD3B1 and CYB5B genes found in these computational analyses.
Bibliography


Appendix A

Supplementary Information for Doubly Sparse Regression Incorporating Graphical Structure Among Predictors

Definitions

**Definition S.1.** An optimal decomposition is any decomposition that achieves the infimum:

\[
\inf_{\mathcal{V}} \sum_{i=1}^{p} \left[ \tau_i \|V^{(i)}\|_2 + \xi \|V^{(i)}\|_1 \right] \quad \text{s.t.} \sum_{i=1}^{p} V^{(i)} = \beta,
\]

for \( \xi > 0, \tau_i > 0 \forall i \), and where \( \mathcal{V} \) is the set of all vectors that may be possible decompositions of \( \beta \) under the support constraint determined by the predictor graph, i.e., \( \mathcal{V} = \left\{ V^{(i)} \in \mathbb{R}^p | V^{(i)}_j = 0 \text{ if } j \notin N_i \right\} \).

**Definition S.2.** (*Negahban et al. (2012)*) The subspace compatibility constant of \( \beta \) and subspace \( \mathcal{M} \) with respect to penalty function \( \mathcal{R}(\cdot) \) and norm \( \| \cdot \| \) can be defined as:

\[
\psi(\mathcal{M}) = \sup \left\{ \frac{\mathcal{R}(\beta)}{\| \beta \|} \forall \beta \in \mathcal{M} \setminus \{0\} \right\}. \quad \text{(S1)}
\]

Proofs and intermediary results

In this section we re-state and prove the estimator properties derived in Section 3.4. Intermediate results are also provided as needed.
Lemma 3.3. Assume (A1)-(A3). Then $\mathcal{R}(\beta)$ is a norm and decomposable with respect to the subspace pair $(\mathcal{M}, \mathcal{M}^\perp)$.

Proof. First note that $\mathcal{R}(\beta) \geq 0$ with equality only when $\beta = 0$. Then for $\phi \in \mathbb{R}\setminus\{0\}$, it is trivial to show positive homogeneity, i.e., $\mathcal{R}(\phi \beta) = |\phi| \mathcal{R}(\beta)$. Lastly, to demonstrate the triangle inequality, let the set of vectors $V^{(i)}, i = 1, \ldots, p$, be a decomposition of the regression parameter vector $\beta$ and $W^{(i)}, i = 1, \ldots, p$, be a decomposition of another regression parameter vector $\beta^*$. Then:

$$
\mathcal{R}(\beta + \beta^*) = \sum_{i=1}^{p} \left[ \tau_i \|V^{(i)}\|_2 + \tau_i \|W^{(i)}\|_2 + \xi \|V^{(i)}\|_1 + \xi \|W^{(i)}\|_1 \right]
\leq \sum_{i=1}^{p} \left[ \tau_i \|V^{(i)}\|_2 + \tau_i \|W^{(i)}\|_2 + \xi \|V^{(i)}\|_1 + \xi \|W^{(i)}\|_1 \right]
= \mathcal{R}(\beta) + \mathcal{R}(\beta^*). \tag{S2}
$$

Thereby proving $\mathcal{R}(\beta)$ is a norm. Then by assumptions (A3), if we have vectors $\beta \in \mathcal{M}$ and $\beta^* \in \mathcal{M}^\perp$ then it follows that $\beta$ and $\beta^*$ will have supports that do not overlap. By (A1) and (A2), it follows that any optimal decomposition of $\beta$, say $V^{(i)}, i = 1, \ldots, p$, and any optimal decomposition of $\beta^*$, say $W^{(i)}, i = 1, \ldots, p$, will also have supports that do not overlap. Since we now have mutually exclusive sets, the triangle inequality (Equation (S2)) will hold with equality and $\mathcal{R}(\beta)$ is decomposable with respect to the subspaces $\mathcal{M}$ and $\mathcal{M}^\perp$.

Lemma 3.4. Suppose $\mathcal{L}(\cdot)$ is a convex and differentiable loss function and consider any optimal solution $\hat{\beta}$ to the optimization problem in Equation (3.3) with a strictly positive regularization parameter satisfying $\lambda \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\beta))$ where $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$ and $\nabla \mathcal{L}(\beta)$ is the gradient of the loss function. Assume (A1)-(A3), and let $\Pi_{\mathcal{M}}(\cdot)$ represent the projection onto the subspace $\mathcal{M}$. Then, the error, $\hat{\Delta} = \hat{\beta} - \beta$, 


will belong to the set:

\[ \mathbb{C}(\mathcal{M}, \mathcal{M}^\perp, \beta) := \{ \Delta \in \mathbb{R}^p | \mathcal{R}(\Pi_{\mathcal{M}^\perp} \Delta) \leq 3\mathcal{R}[\Pi_{\mathcal{M}}(\Delta)] \} \]  

(3.8)

**Proof.** By Lemma 3.3, \( \mathcal{R} \) is a norm and decomposable over \((\mathcal{M}, \mathcal{M}^\perp)\). Negahban et al. (2012) prove that \( \hat{\Delta} \) falls in a star shaped set with upper bound \( 3\mathcal{R}[\Pi_{\mathcal{M}}(\Delta)] + 4\mathcal{R}[\Pi_{\mathcal{M}^\perp}(\beta)] \). By (A3), \( \beta \in \mathcal{M} \) and \( \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\beta)] = 0 \), thus allowing us to omit the last term and \( \mathbb{C}(\mathcal{M}, \mathcal{M}^\perp, \beta) \) will be a cone. \( \Box \)

**Lemma 3.7.** The subspace compatibility constant associated with the optimization in Equation (3.3) is bounded by:

\[ \psi(\mathcal{M}) \leq (\tau_{\text{max}} + \xi \sqrt{d_{\text{max}}}) \sqrt{a}. \]  

(3.10)

**Proof.**

\[
\mathcal{R}(\beta) = \sum_{i=1}^{p} \left( \tau_i \|V^{(i)}\|_2 + \xi \|V^{(i)}\|_1 \right)
\leq \sum_{i=1}^{p} \left( \tau_{\text{max}} \|V^{(i)}\|_2 + \xi \sqrt{d_{\text{max}}} \|V^{(i)}\|_2 \right)
\leq (\tau_{\text{max}} + \xi \sqrt{d_{\text{max}}}) \sum_{i=1}^{p} \|V^{(i)}\|_2
\leq (\tau_{\text{max}} + \xi \sqrt{d_{\text{max}}}) \sqrt{a} \|\beta\|_2. \quad (\text{S3})
\]

The second line of the proof makes use of the property that \( \|V^{(i)}\|_1 \leq \sqrt{(d_{\text{max}})\|V^{(i)}\|_2} \) which will hold by the Cauchy-Schwartz Inequality (Petersen, 2012). The upper bound found in Equation (3.10) can easily be found by substituting the bound in
Equation (S3), into Definition S.2 with respect to the $\ell_2$-norm (i.e., the denominator is $\|\beta\|_2$).

The proof of Lemma 3.8 first requires some intermediary results. We begin by bounding $R^*(\cdot)$, the dual norm of $R(\cdot)$, as follows.

**Proposition S.3.** Assume (A4) and define $u$ to be a $p \times 1$ vector and $u_{V_i}$ to be constrained to have support matching $V(i)$, that is $u_{V_i}$ has non-zero elements $j, j \in V_i$. Then $R^*(u)$ has upper bound:

$$R^*(u) \leq \max_{i=1, \ldots, p} \frac{1}{2\tau_{\min}} \|u_{V_i}\|_2.$$

**Proof.**

$$R^*(u) = \max_{\beta} \left\{ \beta^T u \right\} \text{ s.t. } R(\beta) \leq 1$$

$$= \max_{\beta} \left\{ \sum_{i=1}^{p} V(i)^T u_{V_i} \right\} \text{ s.t. } \sum_{i=1}^{p} \left[ \tau_i \|V(i)\|_2 + \xi \|V(i)\|_1 \right] \leq 1$$

$$\leq \max_{\beta} \left\{ \sum_{i=1}^{p} V(i)^T u_{V_i} \right\} \text{ s.t. } \sum_{i=1}^{p} \left[ 2\tau_i \|V(i)\|_2 \right] \leq 1$$

$$\leq \max_{\beta} \left\{ \sum_{i=1}^{p} V(i)^T u_{V_i} \right\} \text{ s.t. } \sum_{i=1}^{p} \|V(i)\|_2 \leq \frac{1}{2\tau_{\min}} \tag{S4}$$

**Remark:** In deriving Equation (S4), line 3 requires the use of the property that for any vector $a$, $\|a\|_2 \leq \|a\|_1$. By (A4), $\tau_{\max} \leq \xi$. Consequently $\tau_{\max} \|V(i)\|_2 \leq \xi \|V(i)\|_1$ and further $\sum_{i=1}^{p} 2\tau_i \|V(i)\|_2 \leq 1$. Note that $\max_{\beta} \left( \sum_{i=1}^{p} V(i)^T u_{V_i} \right) \leq p \times \max \left( V(i)^T u_{V_i} \right)$ and that the maximum occurs when $V(i)^* = \frac{u_{V_i}}{\frac{1}{2\tau_{\min}} \|u_{V_i}\|_2}$ where $i^* = \arg \max_{i=1, \ldots, p} \|u_{V_i}\|_2$. The result then follows after dividing by $p$ to ensure that $\sum_{i=1}^{p} \|V(i)\|_2 \leq \frac{1}{2\tau_{\min}}$. □

Now an upper bound is required for $R^*(\nabla \mathcal{L}(\beta))$ so that $\lambda$ may subsequently
be bounded.

**Lemma S.4.** Assume (A4) and (A6) and define $\|\nu\|_2^2$ to be a random variable with distribution $\chi_{d_{i_{\max}}}$. Then for the optimization problem in Equation (3.3) $R^*(\nabla L(\beta))$ can be bounded as:

$$R^*(\nabla L(\beta))^2 \leq \frac{\sigma^2\sigma^{*_\max}\|\nu\|_2^2}{4(\tau_{\min})^2 n^2}. \quad (S5)$$

**Proof.** Denote $X_{N_i}$ to be the $n \times d_i$ matrix of predictors that consists of the columns of $X$ in $N_i$, i.e., $X_{N_i} = [X^{(j)}], j \in N_i$. Then, from Proposition S.3 and gradient $\nabla L(\beta)_{N_i} = -\frac{1}{n} X_{N_i}^T \epsilon$, it follows that:

$$R^*(\nabla L(\beta)) \leq \frac{1}{2\tau_{\min}} \max_{i=1,\ldots,p} \|\nabla L(\beta)_{N_i}\|_2 = \frac{1}{2\tau_{\min}n} \max_{i=1,\ldots,p} \|X_{N_i}^T \epsilon\|_2.$$  

Recall $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$; then $X_{N_i}^T \epsilon \sim \sigma \mathcal{N}(0, X_{N_i}^T X_{N_i})$. If $\sigma_i^*$ is defined to be the maximum singular value of $X_{N_i}^T X_{N_i}$ then:

$$\|X_{N_i}^T \epsilon\|_2^2 \leq \sigma^2 \sigma_i^* \|\nu_i\|_2^2,$$

where $\nu_i \sim \mathcal{N}(0, I_{d_i})$. Consequently, $\|\nu_i\|_2^2 \sim \chi_{d_i}^2$ is a chi-square random variable with $d_i$ degrees of freedom. Subsequently:

$$R^*(\nabla L(\beta))^2 \leq \left(\frac{1}{2\tau_{\min}} \max_{i=1,\ldots,p} \|\nabla L(\beta)_{N_i}\|_2\right)^2 \leq \left(\frac{1}{2\tau_{\min}} \max_{i=1,\ldots,p} \| -\frac{1}{n} X_{N_i}^T \epsilon\|_2^2\right)^2 \leq \frac{1}{4 \tau_{\min}^2 n^2} \max_{i=1,\ldots,p} \sigma^2 \sigma_i^* \|\nu_i\|_2^2.$$  

The result in Equation (S5) follows from defining $\sigma^{*_\max} = \max_{i=1,\ldots,p} (\sigma_i^*)$ and setting $\|\nu_i\|_2^2$.
to have the maximum, $d^{\text{max}}$, degrees of freedom.

In Lemma S.6 we use Proposition S.5 (stated without proof in Rao et al. (2013)) to show that the chi-square variable $\| \nu \|^2_2$ defined in Equation (S5) can be bounded with high probability.

**Proposition S.5.** For chi-square random variables $\eta_1, \eta_2, \ldots, \eta_p$ with $d$ degrees of freedom and some constant $c$:

$$
\Pr\left( \max_{i=1,\ldots,p} \eta_i \leq c^2 d \right) \geq 1 - \exp\left( \log(p) - \frac{(c-1)^2 d}{2} \right). \tag{S6}
$$

**Proof.** The Chernoff bound for a random variable $X$ states for every $t > 0$:

$$
\Pr(X \geq \alpha) = \Pr(e^{tX} \geq e^{t\alpha}) \leq \frac{E[e^{tX}]}{e^{t\alpha}} = e^{-t\alpha} E[e^{tX}],
$$

where $E[e^{tX}]$ is the moment generating function for the random variable $X$ (Hoeffding, 1963). Therefore, for $X \sim \chi^2_d$ it follows that $E[e^{tX}] = (1 - 2t)^{-\frac{d}{2}}$ for $t < \frac{1}{2}$. Consequently, the restriction on $t$ becomes $t \in (0, \frac{1}{2})$ and

$$
\Pr(X \geq \alpha) \leq e^{-t\alpha} (1 - 2t)^{-\frac{d}{2}}.
$$

Setting $\alpha = c^2 d$,

$$
\Pr(X \geq c^2 d) \leq e^{-tc^2 d} (1 - 2t)^{-\frac{d}{2}}
$$

$$
= \exp \left\{ -tc^2 d - \frac{d}{2} \log(1 - 2t) \right\}
$$

$$
= \exp \left\{ - \frac{d}{2} \left[ 2tc^2 + \log(1 - 2t) \right] \right\}.
$$
Then let $t = (1 - 1/c^2)/2$ and assume $c > 1$ so that $t \in (0, \frac{1}{2})$. Since we have for all $c > 0$ that $\log(c^2) \leq 2c - 2$ we therefore have,

$$
\mathbb{P}(X \geq c^2d) \leq \exp\left\{-\frac{d}{2} \left[c^2 - 1 - \log(c^2)\right]\right\}
\leq \exp\left\{-\frac{d}{2} \left[c^2 - 1 - (2c - 2)\right]\right\}
= \exp\left\{-\frac{d}{2} \left[c^2 - 2c + 1\right]\right\}
= \exp\left\{-\frac{d}{2} \left[(c - 1)^2\right]\right\}
= \exp\left\{-\frac{(c - 1)^2d}{2}\right\}.
$$

Consequently, considering the union bound for chi-square random variables $\eta_1, \eta_2, \ldots, \eta_p$ each with $d$ degrees of freedom we have:

$$
\mathbb{P} \left( \max_{i=1,\ldots,p} \eta_i \geq c^2d \right) \leq \sum_{i=1}^{p} \exp\left\{-\frac{(c - 1)^2d}{2}\right\}
\leq p \exp\left\{-\frac{(c - 1)^2d}{2}\right\}
= \exp\left\{\log(p) - \frac{(c - 1)^2d}{2}\right\}.
$$

Finally, the result follows by taking the complement to find $\mathbb{P}( \max_{i=1,\ldots,p} \eta_i < c^2d)$. 

Then in Lemma 3.8 we use Lemma S.5 and Proposition S.6 to find a bound for $\mathcal{R}^* (\nabla \mathcal{L}(\beta))$ that holds with high probability.
Lemma 3.8. Assume (A4) and (A6). Then:

$$\mathcal{R}^* (\nabla \mathcal{L}(\beta))^2 \leq \frac{\sigma^2 \sigma_{\text{max}} (\log(p) + d_{\text{max}})}{4 (\tau_{\text{min}})^2 n},$$

(3.11)

with probability at least $1 - c_1 \exp(-c_2 \sqrt{n})$ for some $c_1, c_2 > 0$.

Proof.

$$\mathcal{R}^* (\nabla \mathcal{L}(\beta))^2 \leq \frac{\sigma^2 \sigma_{\text{max}} \|\nu\|^2}{4 (\tau_{\text{min}})^2 n^2} \leq \left( \frac{\sigma^2 \sigma_{\text{max}}}{4 (\tau_{\text{min}})^2} \right) \frac{c^2 d_{\text{max}}}{n} \quad \text{w.p.} \geq 1 - \exp \left( \log(p) - \frac{(c - 1)^2 d_{\text{max}}}{2} \right).$$

(S7)

The result follows from setting $c = r \sqrt{n}$ where $r^2 = \frac{\log(p) + d_{\text{max}}}{d_{\text{max}}}$. Explicit derivations of the constants $c_1$ and $c_2$ follow.

Derivation of Constants for Lemma 3.8

Starting from Equation (S7), we identify $c_1, c_2$ such that the restriction $c_1, c_2 > 0$ holds. From Equation (S7) we have:

$$\mathcal{R}^* (\nabla \mathcal{L}(\beta))^2 \leq \left( \frac{\sigma^2 \sigma_{\text{max}}}{4 (\tau_{\text{min}})^2} \right) \frac{c^2 d_{\text{max}}}{n} \quad \text{w.p.} \geq 1 - \exp \left( \log(p) - \frac{(c - 1)^2 d_{\text{max}}}{2} \right).$$
Then for \( c = r \sqrt{n} \) and \( r^2 = \frac{\log(p) + d_{\text{max}}}{d_{\text{max}}} \), we have \( c = r \sqrt{n} > 1 \) as required. Substituting \( c^2 = r^2 n \) into the Proposition S.5 it follows that:

\[
\mathbb{P}\left( \max_{i=1, \ldots, p} \alpha_i \leq r^2 n d_{\text{max}} \right) \geq 1 - \exp\left( \log(p) - \frac{(r \sqrt{n} - 1)^2 d_{\text{max}}}{2} \right)
\]

\[
= 1 - \exp\left( \log(p) - \frac{d_{\text{max}}}{2} [r^2 n - 2r \sqrt{n} + 1] \right)
\]

\[
= 1 - \exp\left( \log(p) - \frac{d_{\text{max}}}{2} - \frac{d_{\text{max}}}{2} [r^2 n - 2r \sqrt{n}] \right)
\]

\[
= 1 - \left( p \exp\left( -\frac{d_{\text{max}}}{2} \right) \right) \exp\left( - \left( \frac{d_{\text{max}}}{2} [r^2 \sqrt{n} - 2r] \right) \sqrt{n} \right)
\]

\[
= 1 - c_1 \exp\left( -c_2 \sqrt{n} \right)
\]

with

\[
c_1 = p \exp\left( -\frac{d_{\text{max}}}{2} \right) \quad \text{and} \quad c_2 = \left( \frac{d_{\text{max}}}{2} [r^2 \sqrt{n} - 2r] \right).
\]

It is clear that \( c_1 > 0 \). However, for \( c_2 \), since \( r > 1 \) we have \( r^2 > r \) and for \( \sqrt{n} > 2 \) and consequently \( n > 4 \) then \( c_2 > 0 \). We claim this will be valid in any reasonable application of DSRIG.

**BIBLIOGRAPHY**


Appendix B

Additional Results to Chapter 5

The regression coefficients for the simulation study presented in Chapter 5, were taken to be the median regression coefficient from a set of 90 cross-validated DSRIG parameter estimates from a previous study where androstenone was treated as a continuous outcome (Stephenson et al., 2018b). Sparsity among the true regression parameters was enforced by setting the simulation regression parameters to zero if the first or third quartile of the cross-validated estimates were zero. Finally, the simulation regression parameter vector was multiplied by 3. This last step of multiplying by 3 was done to increase the separation of the estimated probability of an outcome of interest. When the simulation regression parameters were left on their original scale, the estimated probabilities for an outcome of interest tended to be concentrated around 0.5 (see Figure B.1(a)). By multiplying the simulation regression parameters by 3, separation was increased (see Figure B.1(b)) and prediction accuracy for the outcome improved.

This Appendix presents additional simulation results to Chapter 5. The simulation study parameters, structure and calculated statistics are the same as those used in Chapter 5 with the following differences: the simulation regression parameters were left on their original scale (i.e., not multiplied by 3), the sample sizes for the training and validation sets were restricted to $\{200, 400, 600, 800, 1000\}$ and only 50 data sets were generated for each sample size.
Figure B.1: Histogram of estimated probabilities for an outcome of interest for regression coefficients multiplied by 1 in (a) and 3 in (b).

Figure B.2: (a) Median proportion of the true non-zero regression coefficients correctly estimated to be non-zero; (b) Overall match rate, given by the median proportion of correctly classified coefficients (zero or non-zero). Proportions based on 50 iterations for five different sample sizes and regression parameters on original scale.
Figure B.3: Median mean absolute rank difference of the absolute value of the coefficients over the true non-zero regression coefficients in (a); and all coefficients in (b). Medians calculated over 50 iterations for five different sample sizes and regression parameters on original scale.
Table B.1: Median proportion of observations misclassified in the test set across 50 iterations for five different sample size of training and validation sets for regression parameters on original scale.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Logistic Regression</th>
<th>min $|\hat{\beta}|_2$</th>
<th>max $|\hat{\beta}|_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LASSO</td>
<td>SLRIG</td>
</tr>
<tr>
<td>200</td>
<td>0.450</td>
<td>0.433</td>
<td>0.431</td>
</tr>
<tr>
<td>400</td>
<td>0.432</td>
<td>0.418</td>
<td>0.416</td>
</tr>
<tr>
<td>600</td>
<td>0.420</td>
<td>0.406</td>
<td>0.404</td>
</tr>
<tr>
<td>800</td>
<td>0.414</td>
<td>0.404</td>
<td>0.403</td>
</tr>
<tr>
<td>1000</td>
<td>0.408</td>
<td>0.400</td>
<td>0.399</td>
</tr>
</tbody>
</table>

Table B.1 records the median proportion of observations misclassified across the 50 iterations of the simulation study by sample size. Similar to results presented in Chapter 5, the misclassification rates for the regularized models are smaller than that of a conventional logistic regression model, particularly when the sample size was small. The incorporation of the graphical structure of the predictors in the SLRIG and DSLRIG models led to a further improvement in prediction accuracy over the LASSO model.

Figure B.2(a) plots the median proportion of the true non-zero regression coefficients correctly estimated to be non-zero, while Figure B.2(b) plots the median proportion of correctly estimated coefficients (zero or non-zero). The SLRIG model tended to identify the largest proportion of the true non-zero regression coefficients but had a poor overall match when compared to the other methods. The overall match rate of the LASSO model was greater than for SLRIG, but it failed to identify a large proportion of the true non-zero regression coefficients. DSLRIG outperformed the LASSO in identification of the true non-zero regression coefficients and had a competitive overall match rate.

Figure B.3(a) plots the median mean absolute rank difference of the absolute value of the coefficients computed over the set of true non-zero regression coefficients,
with the analogous plot computed over all coefficients in Figure B.3(b). For the small sample sizes there is little difference among the candidate models in terms of rank difference over the non-zero set of coefficients. However, SLRIG tends to have the smallest rank difference for large samples sizes followed by DSLRIG and then the LASSO. When computed over all coefficients, the LASSO model using the min $\|\hat{\beta}\|_2$ for tied models tended to have the smallest median rank difference for small sample sizes but as the sample sizes increased, DSLRIG tended to have the smallest median rank difference.
Appendix C

Comparison of Boar Taint Analyses
Figure C.1: Boxplots of cross-validated regression parameter estimates across 65 SNPs for prediction of androstenone expression in the Duroc breed; upper panel corresponds to DSRIG estimates for androstenone treated as quantitative; lower panel corresponds to DSLRIG-max∥\( \hat{\beta} \)∥_2 estimates for dichotomized androstenone. The SNPs are ordered the same in both plots and a similar set of SNPs are associated with large median regression coefficient estimates and boxplots bounded away from zero. SNPs in the HSD3B1 gene and its promoter are found in boxplots 1 through 3 (in blue box) and SNPs in the CYB5B gene are found in boxplots 6 and 7 (in red box).