5-19-2017

High-Dimensional Linear and Functional Analysis of Multivariate Grapevine Data

Uday Kant Jha
ukj7695@rit.edu

ROCHESTER INSTITUTE OF TECHNOLOGY

High-Dimensional Linear and Functional Analysis of Multivariate Grapevine Data

Author:

Uday Kant Jha

Supervisor:

Dr. Peter Bajorski

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

in the

School of Mathematical Sciences

College of Science

May 19, 2017

Declaration of Authorship

I, Uday Kant Jha, declare that this thesis titled, “High-Dimensional Linear and Functional Analysis Of Multivariate Grapevine Data” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Except for such quotations, this thesis is entirely my work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

CERTIFICATE OF APPROVAL

The thesis titled "High-Dimensional Linear and Functional Analysis of Multivariate Grapevine Data" by Uday Kant Jha, a candidate for the degree of Master of Science in Applied Statistics has been examined and approved as worthy of acceptance.

Dr. Peter Bajorski, Professor, School of Mathematical Sciences Date

Thesis Advisor

Dr. Jan van Aardt, Professor, Center for Imaging Science Date

Committee Member

Dr.Ernest Fokoué, Associate Professor, School of Mathematical Sciences Date

Committee Member

ROCHESTER INSTITUTE OF TECHNOLOGY

Abstract

By Uday Kant Jha

School of Mathematical Sciences

Master of Science in Applied Statistics

Variable selection plays a major role in multivariate high-dimensional statistical modeling. Hence, we need to select a consistent model, which avoids overfitting in prediction, enhances model interpretability and identifies relevant variables. We explore various continuous, nearly unbiased, sparse and accurate technique of linear model using coefficients paths like penalized maximum likelihood and nonconvex penalties, and iterative Sure Independence Screening (SIS). The convex penalized (pseudo-) likelihood approach based on the elastic net uses a mixture of the ℓ_1 (Lasso) and ℓ_2 (ridge regression) simultaneously achieve automatic variable selection, continuous shrinkage, and selection of the groups of correlated variables. Variable selection using coefficients paths for minimax concave penalty (MCP), starts applying penalization at the same rate as Lasso, and then smoothly relaxes the rate down to zero as the absolute value of the coefficient increases. The sure screening method is based on correlation learning, which computes component wise estimators using AIC for tuning the regularization parameter of the penalized likelihood Lasso. To reflect the eternal nature of spectral data, we use the Functional Data approach by approximating the finite linear combination of basis functions using B-splines. MCP, SIS and Functional regression are based on the intuition that the predictors are independent. However, high-dimensional grapevine dataset suffers from ill-conditioning of the covariance matrix due to multicollinearity. Under collinearity, the Elastic-Net Regularization path via Coordinate Descent yields the best result to control the sparsity of the model and cross-validation to reduce bias in variable selection. Iterative stepwise multiple linear regression reduces complexity and enhances the predictability of the model by selecting only significant predictors.

Keywords: [Variable Selection; Elastic-Net; Minimax Concave Penalty; Sure Independence Screening; Functional Data Analysis]

Acknowledgements

I would first like to thank my thesis advisor Dr. Peter Bajorski for his tremendous support and help. I learned a lot from him, and this thesis would not have been materialized without his encouragement.

I wish to thank Dr. Jan van Aardt for his acceptance of being on my thesis committee and providing me with the data used in this thesis.

I would also like to thank Dr. Ernest Fokoué for their acceptance of being on my thesis committee and helping with a solution to my queries.

I would also like to thank Grant W.F. Anderson for providing me with the data used in this thesis.

Finally, I would like to thank my parents, siblings, and family for their support and endless love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Due to changing consumption patterns, Technavio analysts forecast the global consumption of wine to reach more than 30 billion liters by 2020. To meet such a huge demand, the study of vineyard leaf spectra becomes the key determinant of grape characteristics like fruit ripening rate, water status, infestation, and disease. Macronutrients including nitrogen, phosphorus, potassium, and magnesium, and the micronutrients including boron, and zinc are found in the soil ("Mineral nutrients," 1998). By studying the leaf biochemistry, we can estimate the nutritional deficiencies caused by micro and macro elements (Zarco-Tejada et al., 2005). According to G. W. Anderson (2016) and Anderson et al. (2016), the six vital nutrients that interest the viticulturists for the growth of wine grapes are nitrogen, potassium, phosphorous, magnesium, zinc, and boron.

Mineral Nutrition and Suppression of Plant Disease, (2014) and Mineral nutrients (1998) explains several essential macronutrients and micronutrients are found in grape vines. Correct amounts of nitrogen (N), as nitrate or ammonium, is necessary for the faster growth of the plant and enhanced rate of photosynthesis. Excessive nitrogen may lead plant to lack resistance to disease whereas its deficiency may cause underdevelopment of the plants and their leaves turn yellow prematurely. Potassium (K) improves root growth, water, and nutrient uptake, and affect the occurrence of a plant disease. Phosphorus (P) plays a vital role in reproduction and metabolism of the plant. Its deficiency may lead to delayed flowering, spindly appearance and bronze-violet pigmentation of leaves and stalks. Magnesium (Mg) is a constituent of chlorophyll. Due to its deficiency, the leaves turn yellow or brown and may shed prematurely. Zinc (Zn) is responsible for fruit set (flowers becoming berries); shoot elongation, pollen development, and antibiotic production to protect the plant cells. Boron (B) is essential for growth and metabolic processes that control plant defense. Its deficiency may reduce the yield of the vines according to G. W. Anderson (2016) and Anderson et al. (2016).

To ensure good crop quality and yield, we need to control the concentrations of these nutrients in plants. The reflectance value of leaf is expressed between 350 – 2500 nm. Hence using electromagnetic reflectance as the input, we can predict the chemical characteristics of these nutrients of grapevine leaves and petioles (Ordóñez, Rodríguez-Pérez, Moreira, & Sanz, 2013).

1.2 Thesis Organization

This thesis has been broadly divided into four parts. The first part, comprising of five chapters (chapters 2 to 6), deals with the theory of all the statistical methodologies used in this thesis. Chapter 2 deliberates about the various aspects of exploratory data analysis, like missing values and outliers, including robust regression and multicollinearity. Chapter 3 discusses various aspects of variable selection and stepwise linear regression. Chapter 4 deliberates about various regularization models using coefficients paths like Ridge, Lasso, Elastic net, Smoothly Clipped Absolute Deviations and Minimax Concave Penalty (MCP). Chapter 5 discusses various aspects of iterative Sure Independence Screening. Chapter 6 deliberates about different aspects of the functional approach to variable selection, including smoothing by basis representation and validation. The second part deals with the reflectance data of the leaves of Riesling, and Cabernet Franc variety of grapes collected from different view angle during their bloom and veraison period of growth. The third part comprising of three chapters (chapter 8 to 10) and deals with the data analysis of one of the various grapevine datasets. In chapter 8, exploratory data analysis is performed on the above-selected data. In chapter 9, selection of variables is carried out using coefficients paths, iterative sure independence screening and functional approach to obtain optimum values of adjusted R–Squared and predicted R–Squared. Then the value of R–Squared, adjusted R–Squared and predicted R–Squared, obtained from various methods, are compared for further study. Chapter 10 explains the problem of dealing with multivariate data. The last part; consist of two chapters (chapter 11 and 12). In chapter 11, four grapevine datasets are chosen from various combinations ensuring representation of each of the two varieties, growth periods, and view angle and analysis of leaf and petiole of the grapevine. Then these datasets are compared based on the best method of variable selection obtained from part three. Chapter 12 provides the conclusion of the thesis.

Chapter 2

Exploratory Data Analysis

2.1 Introduction

The quality of a large real-world data set depends on various issues. The source of the data is the essential factor. Data entry and acquisition is inherently prone to errors, both noncomplex and complex. The field error rates in the data acquisition phase are typically around 5% or more even when the most sophisticated measures to prevent error are used. Recent studies have shown that as much as 40% of the collected data have some or other problem (Maimon & Rokach, 2005). Therefore, for existing data sets the logical solution is to explore the dataset for possible problems and attempt to correct the errors. To enhance the data reliability, data cleansing, such as handling missing values and removal of noise or outliers, becomes necessary. Hence, exploratory data analysis can be regarded as a first step, or a preprocessing step, for any data analysis.

2.2 Missing Values

To find some attribute values missing in much real-life data is ubiquitous in modern research. Missing values is a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. A few missing values on some variables can dramatically shrink the sample size, and if some important attributes missing, then the entire study may fail. There is a variety of reasons why data sets are affected by missing attribute values. Some attribute values were not recorded because they are considered irrelevant, forgotten or placed into the table but later on mistakenly erased. Dealing with missing values requires a careful examination of the data to identify the type and pattern of missing values. Missing data can introduce bias in the parameter estimation. Hence, a suitable method should make that bias as small as possible. The most common approach to handling such missing attribute values is the following method (also called as preprocessing method). The method includes

techniques based on replacing a missing attribute value by the most common value of that attribute, deleting observations with missing attribute values, mean for numerical attributes or value taken from the closest fit case.

2.3 Outliers

In the real datasets, it often happens that some observations, called outliers, are different from the majority. These outliers may be errors, or they could have been recorded under exceptional circumstances, or belong to another population. Hence the first steps towards finding a coherent analysis are the detection of outliers. Although outliers are often considered as an error or noise, they may include relevant information. Detected outliers are candidates for abnormal data that may otherwise adversely lead to model misspecification, biased parameter estimation, and incorrect results. It is, therefore, important to identify them before modeling and analysis. Hawkins defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that a different mechanism generated it. For a field f_i in a record, r_j can be considered as an outlier if the value of $f_i > \mu_i + \varepsilon\, \sigma_i$ or the value of $f_i < \mu_i - \varepsilon\, \sigma_i$. Where μ_i is the mean for the field f_i, σ_i is the standard deviation, and ε is a user defined factor. Regardless of the distribution of the field f_i, most values should be within a certain number ε of standard deviations from the mean. The value of ε can be user-defined, based on some domain or data knowledge (Maimon & Rokach, 2005).

There are numerous modeling methods, which are resistant to outliers or reduce their impact. In the classical least squares (LS) method, which is acutely sensitive to regression outliers, one often tries to detect outliers and replaces them with mean or median. At times, these outliers may contain some useful information; hence, removing or replacing all of them with mean or median may fail to capture the correct pattern. Hence, we need to strike a balance between replacing and retaining outliers.

2.4 Robust Regression

In high dimensional data, the occurrence of outliers is expected, and these outliers may receive considerably more weight leading to distorted estimates of regression coefficients. This distortion makes detection of deviated observations (outliers) difficult because their residuals are much smaller than they would otherwise be without distortion. Also, multivariate leverage outliers can be masked by the effect of good leverage points, on the other hand, some good data points might even appear to be outliers, which is known as swamping. To avoid these effects, robust regression down weights the influence of outliers, which makes their residuals larger and easier to identify. A robust measure is the median of all absolute deviations from the median (MAD):

$$\text{MAD} = 1.483 \underset{i=1,\ldots,n}{\text{median}} |x_i - \underset{i=1,\ldots,n}{\text{median}}(x_i)|$$

A correction factor, constant of 1.483 is used to make the MAD unbiased at the normal distribution. The smallest fraction of outliers called breakdown point that may cause an estimator to take on arbitrarily large aberrant value is around 50% for most of the robust regression method. In other words, robust regression can provide resistant results in the presence of outliers Rousseeuw & Hubert (2011).

2.5 Robust Regression Methods

Linear regression analysis uses the least squares, which would not be appropriate in solving a problem containing outliers or extreme observations. Therefore, we need a parameter estimation method, which is robust where the value of the estimation is not much affected by small deviations in the data. The robust regression applies numerous methods to restrict the influence of outliers; robust regression uses numerous methods. Least Trimmed Squares (LTS) estimation, M-estimation, S-estimation and MM estimation will be explained in robust regression to determine a regression model.

Rousseeuw & Hubert (2011) developed Least Trimmed Squares (LTS) estimation method as given below.

$$\hat{\beta}_{LTS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} (y_i - x_i^T \beta)^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} \epsilon_i^2$$

where $\epsilon_1^2 \leq \epsilon_2^2 \leq \ldots \leq \epsilon_n^2$, are the ordered squared residuals from smallest to largest and *i* = 1, 2, …, *n*. LTS is calculated by minimizing the h ordered squares residuals, where h= [n/2]+[(p+1)/2], with n and p being sample size and number of parameters, respectively. The largest squared residuals are excluded from the summation in this method, which allows those outlier data points to be excluded completely.

The most common method of robust regression is M-estimation, introduced by Huber. Here M indicates an estimation of the maximum likelihood type (Alma, 2011). The M-estimation principle is to minimize the sum of a chosen function ρ of the errors, rather than minimizing the sum of squared errors. The M-estimate objective function is,

$$\hat{\beta}_M = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho \frac{\left(y_i - x_i^T \hat{\beta}\right)}{\hat{\sigma}}$$

where ρ is a symmetric function and continuously differentiable with a unique minimum at zero and $\hat{\sigma}$ is an estimator. An estimate of $\hat{\sigma}$ is given by

$$\hat{\sigma} = \frac{median|\epsilon_i - median(\epsilon_i)|}{0.6745}$$

Iteratively reweighted least squares (IRLS) is used in the calculation of M-estimates. In IRLS, the first fit is calculated, and then a new set of weights is computed based on the results of the original fit. The iterations are continued until a specified number of iterations are finished, or a convergence criterion is met. Thus, function ρ gives the contribution of each residual. However, M-estimation lacks the consideration of the data distribution and uses only the median as the weighted value; hence, it is not a function of the overall data.

To overcome the weaknesses of media, Rousseeuw & Hubert (2011) introduced a high breakdown value method, called S-estimation. Here S indicates that it is based on estimates of scale. S-estimators minimize the dispersion of the residuals, in the same way, that the least squares estimator minimizes the variance of the residuals. Since the S estimate satisfies the necessary conditions as the M estimate, hence it has the same asymptotic covariance as M estimate. The

objective function is minimized residual standard deviation $\hat{\sigma}_s$ ($\in_1(\beta),...,\in_n(\beta)$), where $\in_i(\beta)$ is the i^{th} error term dependent on the regression coefficients β.

$$\hat{\beta}_S = \ min \sum_{i=1}^{n} \rho \frac{\left(y_i - x_i^T \hat{\beta}\right)}{\hat{\sigma}_s}$$

where ρ is a symmetric function and continuously differentiable with a unique minimum at zero and $\hat{\sigma}_s$ is a robust scale estimator. An estimate of $\hat{\sigma}_s$ is given by

$$\hat{\sigma}_s = \frac{median|\epsilon_i - median(\epsilon_i)|}{0.6745}$$

MM estimation is a particular type of M-estimation with an aim to obtain estimates that have a high breakdown value and more efficient. Yohai (1987) developed the MM-estimates as a three-stage procedure. In the first stage, an initial regression parameter is computed using S-estimator, which is consistent, and robust with high breakdown point, but not necessarily efficient.

In the second stage, a more efficient M-estimate of the errors scale is computed using residuals based on the initial estimate. The objective function used in this phase is labeled ρ_0.

Finally, in the third stage, an M-estimate of the regression parameters based on a proper redescending the Psi-function is computed. The last step computes the MM estimate of scale as the solution to

$$\hat{\beta}_{MM} = \text{argmin}_{\beta} \sum_{i=1}^{n} \rho \left(\frac{y_i - x_i^T \hat{\beta}}{\hat{\sigma}_{MM}} \right)$$

where $\hat{\sigma}_{MM}$ is the standard deviation obtained from the residual of S estimation.

2.6 Multicollinearity

In multiple regression models, multicollinearity (also collinearity) refers to a phenomenon in which two or more predictors are highly correlated with each other or the response variable. It increases the variance of the coefficient estimates and makes the estimates very sensitive to minor changes in the model. As a result, the coefficient estimates are unstable and difficult to interpret.

In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant.

To fit the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \in$. The LS solution $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ would usually be sought. However, if $\mathbf{X}^T\mathbf{X}$ is singular, we cannot perform the inversion and the normal equations will not have a unique solution. In this situation, at least one column of $\mathbf{X}$ is linearly dependent on the other columns (i.e., linear combination of the columns of the $\mathbf{X}$ matrix is zero). We would assume "multicollinearity" when there exists a "near dependence" in the $\mathbf{X}$ columns (Draper & Smith, 1998). Multicollinearity can be reduced by removing one of the correlated predictors from the model, because they supply redundant information.

In addition to removing correlated predictors, multicollinearity can be dealt by using other methods, like an elastic net and functional data analysis. The elastic net can select clusters of correlated features when the groups are not known in advance by inducing a grouping or clustering effect during variable selection. These groups of highly correlated variables tend to have coefficients of similar magnitude.

2.7 Variable Selection

Variable selection in multivariate analysis is a critical step in regression, especially when the number of covariates is large in comparison to the sample size. It is an essential step because the removal of non-informative variables will produce better prediction results with simpler models. Hence, the selection methods are based on judicious selection of a subset of variables from the original set, which will allow easier interpretation, better prediction, and reduction in the complexity of the model. Penalized likelihood estimation of the coefficients, based on continuous penalty functions, provide an attractive approach to performing variable selection and estimation of regression coefficient by simultaneously identifying a subset of variables that are associated with a response. In the next three chapters, we discuss various continuous, nearly unbiased, sparse and accurate methods of variable selection using coefficients paths like penalized maximum likelihood and nonconvex penalties, and (SIS).methods of based on convex, non-convex, penalty

and their combination. Also, we discuss the iterative Sure Independence Screening and application of functional data analysis for the high dimensional data.

Chapter 3

Methods of Variable Selection

3.1 Introduction

Traditional multivariate data analytical approaches assume the data of large sample size (n) with a few predictors (p). With the amazing development of modern technology, including computing power and storage, higher dimensional (n ‹‹ p) and high-throughput data of vast size and complexity are being produced for contemporary statistical studies. To perform efficient and reliable model selection for such high-dimensional multivariate data can be challenging. Let $X_1, \ldots, X_p$ is the set of predictors, with n observations and Y be the response variable. The problem of variable selection arises when p is enormous and a subset of $X_1, \ldots, X_p$ is thought to contain many redundant variables. In recent years, the study of such dataset with the curse of dimensionality has received great attention from the research community. Ill-conditioning of the variance-covariance matrix for such high-dimensional dataset renders typical multivariate data analysis unattractive (Wu & Müller, 2010). Hence penalized likelihood procedures can provide an attractive approach for variable selection and regression coefficient estimation by simultaneously identifying a subset of predictors that are associated with a response. For example, C_p (Mallows, 1973), AIC (Akaike, 1974) and BIC (Schwarz, 1978) are all motivated from ℓ_0 penalized likelihood regression. ℓ_0 penalty directly penalizes the number of non-zero coefficients in the model and is intuitively suitable for the purpose of variable selection. However, there are two major limitations in this type of penalized likelihood procedure. First, the ℓ_0 penalty is not continuous at the origin point, and hence the resulting estimators are likely to be unstable. Second, ℓ_0 penalized likelihood procedure involves an exhaustive search over all possible models; hence it is computationally infeasible for a large number of potential covariates. Hence, we require a penalized likelihood estimation based on continuous penalty functions, like ℓ_2-norm or ℓ_1-norm or mixed. Ridge regression (Hoerl & Kennard, 1970) as a continuous shrinkage method, achieves its better prediction performance through a bias–variance trade-off by minimizing the residual sum of squares based on the ℓ_2-norm of the coefficients. However, ridge regression cannot yield a

parsimonious model, for it always keeps all the predictors in the model. Hence it is not a suitable technique for an asymptotic setup (p > n). Regularization technique like Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996) based on ℓ_1-norm, can reduce dimensionality, select variables and estimate coefficients simultaneously. However, it becomes unstable when there is collinearity in the dataset. Hence, a regularization technique like an elastic net (Zou & Hastie, 2005), which can reduce dimensionality, selects variables and encourages a grouping effect simultaneously, appears to be a better option.

3.2 Insight into High Dimensionality

A challenge with high dimensionality is that significant predictors can be highly correlated with some unimportant ones, which increases with dimensionality. The maximum spurious correlation also increases with dimensionality. Consider a situation where all the predictor variable $X_1, \ldots, X_p$ is standardized. The distribution of $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$ is spherically symmetric, $\mathbf{X} = (X_1,\ldots,X_p)^T$ and $\boldsymbol{\Sigma} = \text{cov}(\mathbf{X})$. For better understanding, of the difficulties of high dimensionality we, need to separate the effects of the covariance matrix $\boldsymbol{\Sigma}$ and the distribution of $\mathbf{Z}$ (Fan & Lv, 2008).

When dimension *p* is larger than sample size *n*, then the design matrix $\mathbf{X}$ is rectangular, having more columns than rows. Hence, the matrix $\mathbf{X^T X}$ is large and singular. Due to dimensionality, the spurious correlation between a covariate and the response could be large. The unimportant predictors, which are associated with significant ones (predictors), become highly correlated with the response variable, and the population covariance matrix $\boldsymbol{\Sigma}$ may become ill conditioned as *n* grows. The minimum non-zero absolute coefficient $|\beta_i|$ may decay with *n* and fall close to the noise level.

3.3 Dimensionality Reduction

The curse of dimensionality is strongly linked with the sparseness of data in a high-dimensional space. Dimension reduction or variable selection is an effective strategy to deal with high dimensionality. With dimensionality reduction from high to low, the computational workload

can be radically reduced. Now, accurate coefficient estimation can be found by using one of the well-developed lower dimensional models. The motivation for dimensionality reduction from the original variables is to find the wavelengths significantly responsible for the calculation of various nutrients, rather than linear combinations of all the wavelengths.

We consider the high-dimensional setting of a linear model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

Where dimension of matrix $\mathbf{X}$ is $n \times p$, regression vector $\boldsymbol{\beta}$, $p \times 1$ and response vector $\boldsymbol{Y}$ and ε with $n \times 1$. We denote the active set of variables by

$$S_0 = \{j; \boldsymbol{\beta}_j \neq 0, j = 1, \ldots, p\}$$

The idealistic goal is to make dimensionality reduction with an estimated sparse set of variables

$\hat{S} \subseteq \{1, \ldots, p\}$ such that

$$|\hat{S}| < n\text{-}1$$

Since the data are not high dimensional anymore, one can rely on more classical techniques such as least squares estimation for further analysis using variables from the sparse set $\hat{S}$.

Based on the principle of parsimony, one needs to reduce the amount of complexity in the model while dealing with huge numbers of predictors. To select useful subsets of variables, which may be contributing significantly to the model, we use stepwise regression based on P-values of interest.

3.4 Variable selection

Effective variable selection can lead to parsimonious models with better prediction accuracy and easier interpretation. Ideally, the variable selection procedure should be unbiased, sparse and continuous.

$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$ is modelled by a linear function, where y_i is the response variable, i = 1,…, n, x_{ij} is the explanatory variables, j = 1, . . . ,p, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ are error terms and β_j's are regression, coefficients.

Without loss of generality, we can standardize the response and each covariate with zero mean and unit standard deviation. Hence, after removing the intercept term, the regression model mentioned above can be rewritten as given below.

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

With a large number of predictive variables, we often would like to determine a smaller subset that exhibits the strongest effects. For the purpose of feature selection, we consider the penalized least squares (LS) estimation.

$$= \min_{\beta_j} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} p_\lambda \left(|\beta_j| \right) \right\}$$

where λ is a non-negative tuning parameter and $p_\lambda(\cdot)$ is a sparsity-induced penalty function that which may not depend on λ (Geng, 2014).

The standard techniques for improving the Ordinary Least Square (OLS) estimates are best subset selection, ridge regression, lasso and elastic net. Best subset selection provides models that can be extremely variable because it is a discrete process either retains or drops variables from the model. Its prediction is highly sensitive to minor changes in the dataset. Best subset selection fails, when we have many variables because of several combinations. Ridge regression is a continuous process that improves prediction error by shrinking large regression coefficients to reduce overfitting. However, it fails to perform covariate selection and hence is not very useful when the number of explanatory variables exceeds the number of observation. Because of the l_1 -penalty, Lasso does variable selection and shrinkage, thus retaining the useful techniques of ridge regression and subset selection (Tibshirani, 1996). Hence, it is very helpful when the number of covariates exceeds the sample size. However, it becomes unstable when there is collinearity in the

dataset. To remedy this limitation, one can use Elastic-net regularization, which adds ridge regression-like penalty. It allows the model to select strongly correlated variables together and improves overall prediction accuracy when the number of a covariates is larger than the sample size.

3.5 Stepwise multiple linear regression

In the algorithm for stepwise multiple linear regression, original variables are selected iteratively according to their correlation with the target property. For a selected variable, a regression coefficient is determined and tested for significance using a t-test at a critical level (e.g., 5%). If the coefficient is found to be significant, the variable is retained, and another variable is selected according to its partial correlation with the residuals that are obtained from the model built with the first variable. This procedure is called forward selection. The significance of the two regression coefficients and their association with the two retained variables are then tested again, and the non-significant terms are eliminated from the equation (backward elimination). Forward selection and backward elimination are alternated and repeated until no significant improvement of the model fit can be achieved by including more variables and all regression terms that are already selected are important (Balabin & Smirnov, 2011). In this method, each variable is studied independently, and no consideration is given to variable interaction. The stepwise subset selection approach increases the search space to enhance the predictability of the models. Hence it suffers from statistical problems when p is large and fails in asymptotic setup ($p > n$).

Chapter 4

The regularization models

4.1 Introduction

To reduce variability and achieve a more interpretable model, we often seek a smaller subset of relevant variables. However, searching through subsets of potential predictor variables for an adequate smaller model can be unstable and is computationally unfeasible even of modest dimensions. The objective of variable selection is to identify features in the dataset that are important and discard variables with irrelevant and redundant information. Since variable selection reduces the dimensionality of the data, it holds out the possibility of more efficient & rapid operation of the data set. The ordinary least squares (OLS) estimates often have low bias but significant variance. With great number predictors, we often would like to determine a smaller subset that exhibits the strongest effects. With sparsity, feature selection can improve the accuracy of estimation by effectively identifying the subset of significant predictors, and enhance model interpretability with parsimonious representation. Consider a linear model with a response variable Y, depended on p explanatory variable $X \in \mathbb{R}^p$. For small p, proper penalty on the number of selected variables based on the Cp, AIC, BIC or a data driven method for subset selection can be used to obtain a good guess of the pattern. However, for large p, subset selection is not computationally feasible, so we will have to use continuous penalized or gradient threshold methods.

We describe here algorithms for estimation of linear models with convex penalties, including ℓ_1 (the Lasso), ℓ_2 (ridge regression) and combinations of the two (the elastic net). This algorithm optimizes each parameter separately, holding all other fixed. Updates are trivial. Then it cycles around until coefficients stabilize. This process, called cyclical coordinate descent along a regularization path, achieves dramatic speedups over other competitors. The methods can handle high dimensional data and can deal efficiently with sparse features.

The basic linear regression model used to predict the nutrients with the regularization models is:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{Y}=(y_1,...,y_n)^T$ is the vector of observed response variable, $\mathbf{X}$ is n × p matrix of predictors; $\boldsymbol{\beta}$ is the vector of the regression coefficients of the predictors and ϵ is the vector of the residual errors with variance $(\boldsymbol{\epsilon}) = \sigma_\epsilon^2$. For simplicity, we, assume that the observed variables have been mean-centered, so that we have no need for a constant term in the regression. In other words X_{ij} are standardized, such that $\sum_{i=1}^{n} X_{ij} = 0$, $\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1$, for $j = 1, \ldots, p$.

4.2 Ridge regression

The least square estimate suffers from the deficiency of mathematical optimization techniques that give point estimates. To control the inflation and general instability associated with the least square estimates, one can use ridge regression (Hoerl & Kennard, 1970). Ridge regression performs well only when there is a subset of true coefficients that are small or zero. Ridge regression shrinks all coefficients by a uniform (ℓ_2 – norm) penalty to produce a unique solution. In the case of *k* identical predictors, they each get equal coefficients with $1/k^{th}$ the size, which any single predictor would get if fit alone. Ridge regression is like least squares but shrinks the estimated coefficients towards zero. For a given response vector $\mathbf{Y} \in \mathbb{R}^n$ and a predictor matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, the ridge regression coefficients are defined as:

$$\hat{\beta}^{(\text{ridge})} = \underset{\beta \epsilon \mathbb{R}^p}{\arg\min}\{\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_2^2\}$$

Where $\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$ is the ℓ_2 –norm (quadratic) loss function (i.e. residual sum of squares), x_i^T is the i^{th} row of $\mathbf{X}$, $\| \boldsymbol{\beta} \|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the ℓ_2 –norm penalty on $\boldsymbol{\beta}$, and $\lambda \geq 0$ is the tuning (penalty, regularization, or complexity) parameter which regulates the strength of the penalty (linear shrinkage) by determining the relative importance of the data-dependent empirical error and the penalty term. As λ tends to infinity, the coefficients will approach zero. In other words, the larger the value of λ, the greater is the amount of shrinkage. As the value of λ is dependent on the data, it can be determined using data-driven methods, such as cross-validation.

The intercept is assumed to be zero due to mean centering of the variables. (Schulz-Streeck, Ogutu, & Piepho, 2012). Since ridge regression does not set the coefficients exactly to zero unless $\lambda = \infty$, in which case all the coefficients are zero. Hence ridge regression cannot select a model with the most relevant and predictive subset of predictors.

4.3 Least Absolute Shrinkage and Selection Operator (Lasso)

Due to the nature of the ℓ_1-penalty, the lasso does both continuous shrinkage and automatic variable selection simultaneously. Even though the prediction performance of the Lasso and Ridge regression are similar, however, the Lasso is more appealing due to its sparse representation (Zou & Hastie, 2005). The Lasso shrinks the magnitude of all the coefficients by a constant value and sets them to zero if they reach that value, as in the best subset selection case. In other words, Ridge regression shrinks all regression coefficients towards zero; the Lasso tends to give a set of zero regression coefficients, which leads to a sparse solution. The Lasso penalty corresponds to a Laplace prior, which expects a large number of coefficients to be zero, and only a small subset to be nonzero. The Lasso estimator uses the $\ell 1$ penalized least squares criterion to obtain a sparse solution to the following optimization problem:

$$\hat{\beta}^{(\text{Lasso})} = \underset{\beta \epsilon \mathbb{R}^p}{\arg\min}\{\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1\}$$

Where $\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 = \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$ is the ℓ_2 –norm (quadratic) loss function (i.e. residual the sum of squares), x_i^T is the i[th] row of $\mathbf{X}$, $\| \boldsymbol{\beta} \|_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 –norm penalty on $\boldsymbol{\beta}$, which induces sparsity in the solution, and $\lambda \geqslant 0$ is the tuning parameter.

The ℓ_1 penalty enables the Lasso to simultaneously regularize the least squares fit and shrink some components of $\hat{\beta}^{\text{Lasso}}$ To zero for some suitably chosen λ (Schulz-Streeck et al., 2012). Although the Lasso has many excellent properties, it is a biased estimator and the bias for a truly nonzero variable is about λ for large regression coefficients. It is not robust to highly correlated predictors. It also fails to do grouped selection. If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to arbitrarily pick one variable from the group and ignore the others. In the extreme case when all predictors are identical, the lasso breaks

down. In addition, Lasso also restricts the number of variables that can be selected. If p>n, the lasso selects at most n variables.

4.4 Elastic net

From the Bayesian standpoint, the ridge penalty is ideal when there are many predictor variables with non-zero coefficients (drawn from a Gaussian distribution). The Lasso penalty, on the other hand, corresponds to a Laplace prior that assumes many coefficients to be zero, and only a small subset to be nonzero. The elastic net is a compromise between ridge and lasso that is robust to extreme correlations among the predictors. The elastic net encourages a grouping or clustering effect when strongly correlated predictors enter or exit the model together. The elastic net is mainly useful when the number of predictors (p) is bigger than the sample size (n). The elastic net is very helpful to analyze high dimensional data and to avoid the instability of the lasso solution paths when pairwise correlations are very high. The elastic net uses a mixture of the ℓ_1 (lasso) and ℓ_2 (ridge regression) penalties. It is formulated as given below:

$$\hat{\beta}^{(\text{enet})} = \arg\min_{\beta \epsilon \mathbb{R}^p} \{ \| \mathbf{Y} - \mathbf{X\beta} \|_2^2 + \lambda P_\alpha(\beta) \}$$

where $\| \mathbf{Y} - \mathbf{X\beta} \|_2^2 = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$, and $P_\alpha(\beta)$ is the elastic net penalty subject to

$$P_\alpha(\beta) = (1-\alpha) \| \boldsymbol{\beta} \|_2^2 + \alpha \| \boldsymbol{\beta} \|_1 = \sum_{j=1}^{p} [\tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|] \leq s \quad \text{for some s}$$

$P(\beta)$ creates a useful compromise between the ridge-regression penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). The ℓ_1 part of the elastic net does automatic variable selection, while the ℓ_2 part encourages grouped selection and stabilizes the solution paths with respect to random sampling, thereby improving prediction. As α increases from zero (0) to one (1), for a given λ the sparsity of the solution (i.e., the number of coefficients equal to zero) increases monotonically from zero to the sparsity of the lasso solution. This penalty is particularly useful in the $p \gg n$ situation, or any situation where there are many correlated predictor variables However, unlike the lasso, when $p \gg n$, the elastic net may select more than 'n' variables (Schulz-Streeck et al., 2012).

4.5 Smoothly Clipped Absolute Deviations (SCAD)

It is known that the ℓ_2 does not satisfy the sparsity condition, and the convex ℓ_1 penalty does not meet the unbiasedness condition and the concave ℓ_q penalty with $0 \leq q < 1$ does not meet the continuity status. In other words, none of these ℓ penalties satisfies all three conditions simultaneously. In high a dimension condition, the bias of penalized estimators can almost be removed by choosing a constant penalty beyond a second threshold level $\gamma\lambda$. Fan & Lv (2010) introduced the Smoothly Clipped Absolute Deviation (SCAD), which retains the penalization rate (and bias) of the lasso for small coefficients, but continuously relaxes the rate of penalization as the absolute value of the coefficient increases. The SCAD penalty is continuously differentiable on (-∞, 0) U (0, ∞), but singular at zero with its derivatives zero outside the range $[-\gamma\lambda, \gamma\lambda]$. These results in small coefficients being set to zero, a few other coefficients being shrunk towards zero while retaining the large coefficients as they are. Thus, SCAD can produce sparse set of solution and approximately unbiased coefficients for large coefficients. Fan & Lv (2008) defined the continuously differentiable penalty SCAD by

$$p'_\lambda(|\beta|) = \lambda\left\{I(|\beta| \leqslant \lambda) + \frac{(\gamma\lambda - |\beta|)_+}{(\gamma-1)\lambda} I(|\beta| > \lambda)\right\} \quad \text{for some } \gamma > 2$$

where $p'_\lambda(|\beta|)$ is a concave penalty with respect to $|\beta|$. The authors suggested using $\gamma = 3.7$. It coincides with the Lasso until $|X| = \lambda$, then smoothly transit to a quadratic function until $|X| = \gamma\lambda$, after which it remains constant for all $|X| > \gamma\lambda$. These results apply to general classes of loss and penalty functions but do not address the uniqueness of the solution or provide methodologies for approximating the local minimizer with the stated properties. A major cause of computational and analytical difficulties in these studies of nearly unbiased selection methods is the non-convexity of the minimization problem.

4.6 Minimax Concave Penalty (MCP)

The Minimax Concave Penalty (MCP) starts by applying the same rate of penalization as the lasso, and then smoothly relaxes the penalization rate to zero as the absolute value of the coefficient increases. The MCP relaxes the penalization rate immediately, whereas for SCAD the rate remains flat for a while, before decreasing.

Zhang (2010) defined MCP as,

$$\rho(t;\lambda) = \lambda \int_0^t \left(1 - \frac{x}{\gamma\lambda}\right)_+ dx,$$

with a regularization parameter $\gamma > 0$. It minimize the maximum concavity

$$\kappa(\rho) \equiv \kappa(\rho;\lambda) \equiv \sup_{0<t_1<t_2} \{\dot{\rho}(t_1;\lambda) - \dot{\rho}(t_2;\lambda)\}/(t_2 - t_1)$$

subject to the following unbiasedness and features selection:

$$\dot{\rho}(t;\lambda) = 0 \quad \forall t \geq \gamma\lambda, \qquad \dot{\rho}(0+;\lambda) = \lambda.$$

Convexity ensures that the algorithm converges to the unique global minimum and $\hat{\beta}$ is continuous with respect to λ, which in turn ensures good initial values, thereby reducing the number of iterations required by the algorithm. In the absence of convexity, $\hat{\beta}$ is not necessarily continuous with respect to the data—that is, a small change in the data may produce a large change in the estimate. Such estimators tend to have high variance in addition to being unattractive from a logical perspective. Besides, discontinuity with respect to λ increases the difficulty of choosing a good value for the regularization parameter. The coordinate descent algorithms are also not guaranteed to converge to a global minimum in general. However, it is not always necessary to attain global convexity. In high-dimensional settings where p > n, global convexity is neither possible nor relevant. In such settings, sparse solutions for which the number of nonzero coefficients is much lower than p, we will still have stable estimates and smooth coefficient paths in the parameter space of interest (Breheny & Huang, 2011).

MCP provides the sparse convexity to the broadest extent by minimizing the maximum concavity. The MCP achieves $\kappa(\rho;\lambda) = 1/\gamma$. A larger value of its regularization parameter γ affords less unbiasedness and more concavity. For each penalty level λ, the MCP provides a continuum of penalties with the ℓ_1 penalty as $\gamma \longrightarrow \infty$ i.e., the MCP and lasso solutions are the same, and the "ℓ_0 penalty" as $\gamma \longrightarrow 0+$ (Zhang, 2010).

Chapter 5

Sure Independence Screening

5.1 Introduction

Variable selection plays a major role in high-dimensional statistical modeling, which nowadays appears in many areas and is key to various scientific discoveries. For problems of high dimensionality p, the accuracy of estimation and computational cost are two top concerns. One popular family of feature selection methods for parametric models is based on the penalized (pseudo-)likelihood approach. It includes the Lasso (Tibshirani, 1996), the SCAD (Fan & Li, 2001), the elastic net (Zou & Hastie, 2005), the MCP (Zhang, 2010), and related techniques. Nevertheless, in ultrahigh dimensional statistical learning problems, these methods may not perform well due to the concurrent challenges of computational expediency, statistical accuracy, and algorithmic stability. Motivated by these concerns, Fan & Lv (2008) introduced the concept of sure screening method based on correlation learning, called sure independence screening. It reduces dimensionality from high to a moderate scale, below the sample size. As a methodological extension, iterative sure independence screening is also proposed to enhance its finite sample performance.

5.2 Sure Independence Screening

Consider estimating a p-vector of parameters β from the linear model

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\epsilon,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is an n-vector of responses, $\mathbf{X} = (X_1,..., X_n)^T$ is n x p matrix, which is independent and identically distributed. $X_1,..., X_n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-vector of parameters and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is an n-vector of IID random errors. When the dimension p is high, it is assumed that only a small number of predictor variables among $X_1, \ldots, X_p$ contribute to the

response, which amounts to assuming ideally that the parameter vector β is sparse. With sparsity, variable selection can improve the accuracy of estimation by effectively identifying the subset of important predictors, and enhance model interpretability with parsimonious representation. Sparsity comes frequently with high dimensional data, which is a growing feature in many areas of contemporary statistics. The problems arise frequently in genomics, imaging, and finance, where the number of variables or parameters p are much larger than sample size n. Let us assume that the predictors $X_1, \ldots, X_p$ are independent and follow the standard normal distribution. Then, the design matrix is an n x p random matrix, each entry an independent realization from N (0, 1). The maximum absolute sample correlation coefficient between predictors can be very large. The multiple canonical correlation between two groups of predictors (e.g. 2 in one group and 3 in another) can be even larger. We can filter out the predictors, which have weak correlation with the response using the concept of sure independence screening. By sure screening, Fan & Lv (2008) mean that all the important variables survive after applying a variable screening procedure with probability tending to 1. Fan & Lv (2008) introduces a simple sure screening method using component wise regression or equivalently correlation learning, where input variables are independent and follow the standard normal distribution N (0, 1).

Let $\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$ be the true sparse model with non-sparsity size s = $|\mathcal{M}_*|$. The other (p – s) variables can also be correlated with the response variable via linkage to the predictors that are contained in the model. Let $\boldsymbol{\omega} = (\omega_{1,\ldots,}\omega_p)^{\mathrm{T}}$ be a p-vector that is obtained by component-wise regression, i.e.

$$\boldsymbol{\omega} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

Where n x p data matrix $\mathbf{X}$ is first standardized column-wise. Hence, $\boldsymbol{\omega}$ is a vector of marginal correlations of predictors with the response variable, rescaled by the standard deviation of the response. For any given $\gamma \in (0, 1)$, we sort the p component-wise magnitudes of the vector ω in a decreasing order and define a sub-model

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\},$$

Where $[\gamma n]$ signifies the integer part of γn. This is a straightforward way to shrink the full model $\{1, \ldots, p\}$ down to a sub-model $\mathcal{M}_\gamma$ with size d = $[\gamma n]$ < n. Such correlation learning ranks the

importance of variable according to their marginal correlation with the response variable and filters out those that have weak marginal correlations with the response variable. This correlation screening method is called SIS, since each variable is used independently as a predictor to decide how useful it is for predicting the response variable and the concept is applicable to generalized linear models (Fan & Lv, 2008).

Chapter 6

Functional Data Analysis

6.1 Introduction

In the last few decades, data collection technology has evolved to measure observations densely sampled over time, wavelength, space and other continua. For modeling this type of data, it is more natural to think in functional terms even though only finite numbers of observations are available. In such case, the random variables can take values into an infinite dimensional space and is represented by a set of curves. Theoretically, the infinite dimension is the largest source of difficulty in modeling such data (Jacques & Preda, 2014). Since an observed value is available at each point on a line segment, a portion of a plane, hence, curves and images can be considered as functions. For this reason, we call observed curves as functional data, and statistical methods for analyzing such data are termed functional data analysis (J. O. Ramsay & Dalzell, 1991). In recent years, many researchers have proposed various methods to solve these functional data, including functional regression analysis, functional principal components analysis, functional clustering, and functional multi-dimensional scaling (Mizuta & Kato, 2007). A functional regression model, which is the functional version of the regression model, can provide a useful tool for analyzing such dataset (Matsui, Kawano, & Konishi, 2009).

6.2 Functional Data

A functional datum is not a single observation, but rather a set of measurements along a continuum that, taken together, are regarded as a single entity, curve or image belonging to an infinite dimensional space (Levitin, Nuzzo, Vines, & Ramsay, 2007). Let a functional variable X be a random variable taking values in an infinite dimensional space (or a functional space) E. Then; a technical dataset is just a sample $\{X_1,\ldots, X_n\}$ drawn from a functional variable X. Here, E is assumed to be a normed or semi normed metric space. If functional data are sparsely sampled, or there are many missing data points or E is a Hilbert space, then probably the representation in a basis is mandatory. If a random variable can be observed at different times in the range (t_{min}, t_{max}),

then the observation can be expressed by the random family { $X_1(t),\ldots,X_n(t)$ }. In other words, we can consider the data as an observation of the continuous family $X = \{X(t); t \in T$ (time interval or wavelengths)}. We restrict ourselves to the case where E is a space of real-valued functions (Jacques & Preda, 2014).

6.3 Proximities Notions

Proximities measure between mathematical objects play a major role in all statistical methods. In a finite dimensional Euclidean space ($\mathbb{R}_p$) there is an equivalence between all norms. The most popular in $\mathbb{R}_p$ is the Euclidean norm ||.||, which is based on the sum of squares of the components of any vector.

Let $\boldsymbol{X} = (X_1,..., X_p)^T$ be a vector of $\mathbb{R}_p$; then, the classical Euclidean norm is defined by

$$||\boldsymbol{X}||^2 = \sum_{j=1}^{p}\left(\boldsymbol{X_j}\right)^2 = \boldsymbol{X^T X}$$

However, in an infinite dimensional space, the equivalence between norms fails. In the functional context, the choice of the preliminary norm becomes more crucial, especially when the normed or metric spaces is too restrictive. In such case, semi-metric spaces are better adapted than metric spaces. By definition a semi-norm ||.|| which is similar to norm except that $||\boldsymbol{X}|| = 0 \not\Rightarrow X = 0$. Similarly, a semi-metric d can be defined to be a metric but such that $d(x, y) = 0 \not\Rightarrow x = y$. In other words, the semi-metrics act as a filter and a "good semi-metric" will be a priori, which can select all the pertinent information (Ferraty & Vieu, 2006).

6.4 Functional Regression Model

In the linear regression, both the response variable Y and the predictors (covariates) X_j are scalar, and the model takes the form

$$Y = \Sigma_{j=0}^{p}\beta_j X_j + \epsilon\,, \quad j= 1,2,\ldots, p$$

The error term ∈ allows for sources of variation, such as measurement error, trivial causal factors, and are assumed to be independently and identically distributed. However, this model does not account for the fact that X_1 represents a wavelength that is right next to the wavelength of X_2, and so on. In other words, the above linear model fails to capture the smoothness of the X variables on the wavelength. In such a situation, using a functional approach makes more sense. Functional regression analysis is widely used to describe the relationship between response and predictor variables when at least one of the variables contains a random function. We can convert the data to a functional form in two steps: choose and define a set of basis functions, and compute the best linear combination.

If we replace at least one of the p covariate observations $X_i = (X_{i1, \ldots,} X_{ip})$ in the linear equation by a functional covariate $X_i(t)$, we get a model consisting of a single functional independent variable, plus an intercept term.

Now, we can discretize each of the n functional covariates $X_i(t)$ by choosing a set of times $t_1, \ldots, t_q$ and consider fitting the model.

$$Y_i = \alpha_0 + \Sigma_{j=0}^{q} x_i(t_j)\beta_j + \epsilon_i$$

If we continue refining the selected time, the summation will approach an integral equation, and we will get a functional linear regression model for the scalar response:

$$Y_i = \alpha_0 + \int x_i(t)\beta(t)\, dt + \epsilon_i\,, \qquad i = 1, \ldots, n \;\; Y_i \sim N(\mu, \sigma^2)$$

where the functional regression seeks to quantify the relationship between a scalar outcome Y_i and a random functions $x_i(t)$ (J. O. Ramsay, Hooker, & Graves, 2009).

Here the constant α_0 is the intercept term that adjusts for the origin of the response variable. The parameter β is in the infinitely dimensional space of ℓ_2 functions (the Hilbert space of all square integral functions over a certain interval) (Febrero-Bande & Oviedo de la Fuente, 2012).

6.5 Smoothing by Basis representation

If we consider each time t as index for a separate scalar independent variable, X(t) then the model will look like any conventional multiple regression. However, now we will have

potentially infinite independent variables at our disposal to predict limited number of scalar values, which will result in over-fitting of the data. To avoid this problem, we approximate a function with a finite linear combination of basis functions using B-splines (piece wise polynomial). When we assume the data is d to belong to ℓ_2 space, then we can represent a curve by a basis. A basis is a set of known functions $\{\phi_k\}_{k \in \mathbb{N}}$ that any function could be arbitrarily approximated by taking a weighted sum or a linear combination of a sufficiently large number *K* of these functions (Febrero-Bande & Oviedo de la Fuente, 2012).

Basis function procedures represent a function $X(t)$ by using a fixed truncated basis expansion regarding *K* known basis elements,

$$X(t) = \sum_{k \in \mathbb{N}} c_k \phi_k(t) \approx \sum_{k=1}^{K} c_k \phi_k(t) = c^T \mathbf{\Phi}(\mathbf{t})$$

The smoothing (or hat) matrix H is square, symmetric and of order n.

$$H = \Phi(\Phi^T\Phi)^{-1}\Phi^T,$$

The effective degrees of freedom for functional fit is defined by;

$$df = \text{trace}(H) = K,$$

moreover, the associated degrees of freedom for error is n – df.

When smoothing penalization λ is used then the hat matrix H is given by:

$$H = \Phi(\Phi^T\Phi+\lambda R)^{-1}\Phi^T,$$

where *R* is the penalization matrix, with the integral of the square of the derivative of order 2.

As $\lambda \longrightarrow 0$, $df(\lambda) \longrightarrow \min(n, K)$, where n = the number of observations and *K* = the number of basis functions. Similarly, as $\lambda \longrightarrow \infty$, $df(\lambda) \longrightarrow m$, where m is the order of the highest derivative used to define the roughness penalty.

The regression approach to smoothing data only works if the number *K* of basis functions is substantially smaller than the number of observations. Larger values of K will tend to undersmooth or overfit the data (J. O. Ramsay et al., 2009).

6.6 Validation Criterion

The choice of the smoothing parameter is important and, in principle, no universal rule would enable an optimal choice. Among the different selection criteria to select the parameter λ, we will discuss two: Cross-validation (CV) and generalized cross validation (GCV). The basic idea behind cross-validation is to set part of the data to one side, calling it a validation sample, and fit the model to the balance of the data, called the training sample. In that way, we see how well the model fits data that were not used to estimate the model, thus avoiding the somewhat incestuous procedure of using the data to both fit the model and assess fit. However, the method is not suitable for large sample sizes due to computational intensity, and minimizing CV can lead to under smoothing the data. To overcome these problems, generalized cross-validation criterion (GCV) was developed to locate a best value for smoothing parameter (λ) (J. Ramsay & Silverman, 2005). The criterion is

$$\text{GCV}(\lambda) = \left(\frac{n}{n-df(\lambda)}\right)\left(\frac{SSE}{n-df(\lambda)}\right)$$

The right factor is the unbiased estimate of error variance σ^2 familiar in regression analysis and thus signifies some discounting by subtracting $df(\lambda)$ from n. The left factor further discounts this estimate by multiplying by $n/(n - df(\lambda))$.

Chapter 7

Grapevine Data

7.1 Location

The vineyard farm is located in Lansing, NY (42^{o}34′22.1″ N, 76^{o}35′47.9″ W) with two different grape cultivars, used for data collection: a block of Riesling (shown in blue in Figure -7.1), and a block of Cabernet Franc (shown in red in `Figure 7.1`). Since different soil management treatments were applied to these blocks over the last several years, one could expect a broad range of nutrient concentrations in the leaves. This soil management treatment should add variability to the spectral data, which translates to a range of values across the collected spectral and nutrient data.

`Figure 7.1`: Location of the farm for data collection

7.2 Spectral Data Collection

Reflectance spectra were collected during the bloom and the veraison period of growth for the nutrient analysis from the farm, located in Lansing NY in two separate grape blocks namely Riesling, and Cabernet Franc. For Riesling field, two sets of three panels in each row were grouped together in a single block to give us 24 different samples. The Cabernet Franc field consisted of 8 rows with four viable panels in each. Each panel was considered a block on its own giving the nutrient analysis of 32 unique samples. One of the panels from the first row was dead, resulting in

the reduction of the total number of sample for this grape variety down to 31. These blocks were selected for analysis as they had different soil management treatments applied to them throughout the last several years, which theoretically should have resulted in a wide range of nutrient concentrations in the leaves. The data were then averaged to match that of the cultivar nutrient sampling approach. Data for Cabernet Franc were obtained by averaging the two samples taken from each view angle in each panel and lumped together in the nutrient sampling. On the other hand, data for Riesling involved averaging the six samples collected between the three panels according to G. W. Anderson (2016) and Anderson et al. (2016).

7.3 Nutrient Analysis

The collection of samples from the grape vines were timed such that they were collected, within hours of the spectral samples being collected from the vines, and typically within minutes of the spectra being collected. According to typical nutrient analysis, the petioles for each panel were collected, dried, ground up and combined before analysis. The petioles from the vines in each panel were collected and prepared using the viticulture standard method mentioned by G. W. Anderson (2016) and Anderson et al. (2016). A second nutrient analysis was conducted on the leaf blades, prepared in the same way as the petioles, to compare the results between the petioles and leaves.

7.4 Spectral Reflectance

During the bloom data collection, a traditional the Spectralon (reflectance coefficient = 0.993) panel for calibration was used as it has a near 100% reflectance across the 350- 2500nm range. However, due to non-availability of the Spectralon panel, a section of Tyvek (reflectance coefficient = 0.97) was used for data collection during the veraison season. Since we are concerned with the relative reflectance of the grape leaves, and the difference in their reflectance coefficient does not matter.

Two samples per vineyard panel were selected for data collection to overcome spurious results, using a Spectra Vista Corporation (SVC) spectroradiometer, (SVC HR-768i) for each of three different view angles. Then the collected data were averaged. The first view angle was at nadir for the individual grape leaves by holding the SVC approximately 0.30m (+0.03m/-0.10m) from each leaf. The second view angle was the vine canopy at nadir using a ladder beside the row of grape vines and holding the sensor approximately 1m (+0.3m/-0.3m) above the bulk of the canopy. The third view angle was canopy at 15° off-nadir, using a ladder, with the sensor held approximately 1m (+0.3m/-0.3m) parallel to the side of the row G. W. Anderson (2016) and Anderson et al. (2016).

Chapter 8

Exploratory Data Analysis of Grapevine Data

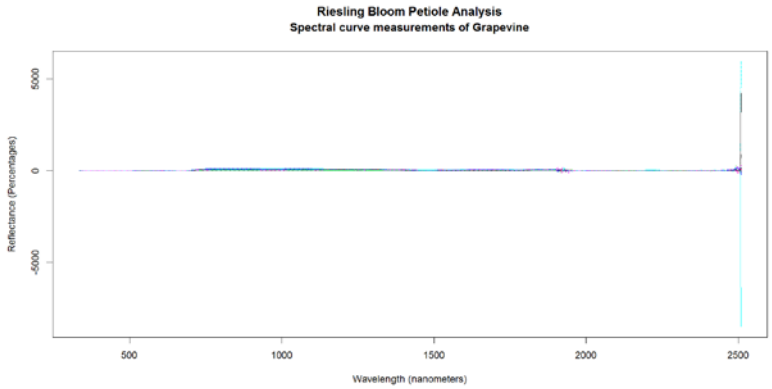
8.1 Data Analysis Methods

The grapevine data were collected for two different grape cultivars, namely Riesling and Cabernet Franc, and growing period, viz. bloom and veraison from three separate angles against the nutrient data. The data, their source, and the data collections efforts are described in G. W. Anderson (2016) and Anderson et al. (2016). Among the various combination of grapevine data, we selected one dataset, about the Petiole Analysis, where the reflectance has been taken directly from the individual leaves of the Riesling variety during the bloom period of growth. The 986 observations of spectral data were collected for the wavelength spread from 334 nanometers (nm) to 2510 nanometers. These spectral data were read in R–studio by merging 144 files containing Spectra Vista SIG data against their wavelength. This information was transposed to obtain one value for each file against the 986 different values of wavelength. Thus a table with 144 rows and 986 variables are formed. Since there was only 24 observation for the six different nutrients, namely nitrogen, potassium, phosphorus, magnesium, zinc, and boron, each value were replicated six times to match the number of rows of the merged file. The data are merged with the spectral data, to obtain a matrix of 144 rows and 987 variables, i.e., 986 predictors and one dependent variable. Since we are interested in detecting and predicting the nutrients values of the grapevine by using reflectance, we treat this as a regression problem. Hence, after merging the files, we need to explore the data for any missing values, outliers, and multicollinearity, before proceeding with prediction.

After scrutinizing the data, we found that there were no missing values for the predictor as well as response variables.

8.2 Outliers

One of the challenges in data analysis is dealing with outliers. When analyzing data, outliers cause problems because they may strongly influence the result. We can check the outliers by plotting the spectral curve measurement of the wavelengths against their value of reflectance as shown in `Figure 8.1`.



`Figure 8.1`: Spectral Curve measurement of the Reflectance against the wavelength

From the `Figure 8.1`, we can see there are some outliers. Since the value of reflectance is given as a percentage, it cannot exceed 100 and at the same time cannot contain negative values. Detailed study shows that there are 828 and 72 observations with values more than 100, often due to atmospheric noise, and less than zero respectively, indicating an error in data collection or entry.

Outliers could not only represent an inaccuracy in the data, but they may also indicate a significant new trend. It might be the clue to data behaviors that are not revealed by the rest of the information. Hence, an optimum balance between replacing and retaining outliers needs to be considered. However, in this case, due to the presence of certain extreme outliers the standard deviation is very high, resulting in Coefficient of Variation of around 125 percent. The standard convention of considering values more than three standard deviations as outliers fail to resolve the issue. Hence, we replace all values less than 0 and more than 100 percent by the mean value of the matrix, which is within one and two standard deviations, respectively. We often omit values near 1400 nm, 1900nm, and 2100nm due to atmospheric noise.

The matrix of predictor variables has 144 observation and 986 covariates with values of reflectance in percentage. In this matrix, there are 900 wrong observations with value more than

100 and less than 0 (zero). We replace these bad observations with a mean value of the matrix. Now we can plot the spectral curve measurement of the wavelength against their value of reflectance in percentage.

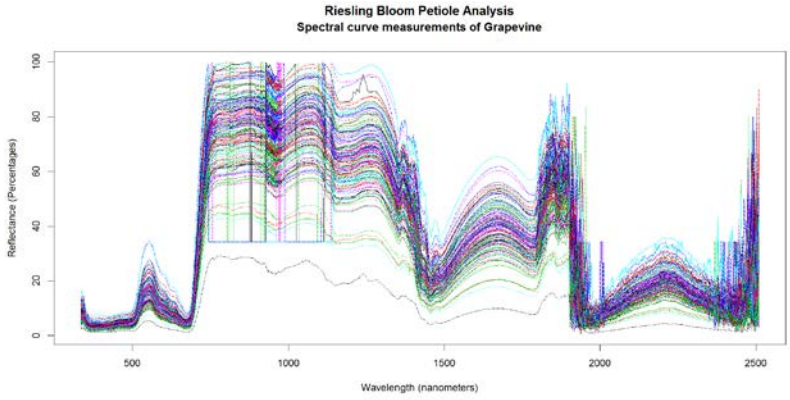


Figure 8.2: Spectral Curve measurement of the Reflectance against the wavelength after replacing w observations with mean

From the Figure 8.2, we can see that certain portions of the spectrum of wavelengths under study indicate the higher values of reflectance in the percentage term than the other portions. In particular, the wavelengths between 700 to 1400 nanometers (near infrared range) and from 1800 to 1900 nanometers (atmospheric noise) shows the higher value of reflectance in percentage term. Again, isolated wavelengths between 2400 and 2500 nanometers due to low signal or noise could be seen with a high value of reflectance. Also, the spectral reflectance is highly correlated to the tune of 98% in certain cases as expected, because they are observed at wavelengths separated by 1.5 to 2.7 nanometers. This multicollinearity increases the standard errors of the coefficients, which in turn indicates that coefficients for some independent variables may not be significantly different from zero.

Since there is no single model to establish the importance of outliers in the given data, we will exam this with the help of the value we obtain for R-Squared, adjusted R-Squared and predicted R-Squared. We will compare three models. In the first linear regression model, we will use the original matrix of 144 observation and 987 covariates. In the second linear regression model, we will replace the 900 wrong observations, from the input matrix of 144 observation and 986 covariates, with values more than 100 and less than 0 (zero) by the mean. In the third model,

we will use the original matrix of 144 observation and 987 covariates but apply the robust regression.

Since the given dataset has more covariates than the sample size, hence the matrix suffers from the curse of dimensionality. Therefore, ordinary least square cannot be performed on this data. On the other hand, ridge regression as a continuous shrinkage method will be able to achieve better prediction performance. However, it cannot produce a parsimonious model, for it always keeps all the predictors in the model. To perform linear regression, we need to reduce the number of covariates from 987 to less than 144. Owing to the nature of the convex optimization, in high dimensional case (p > n), Lasso can select maximum *n* variables before it saturates, by continuous shrinkage and automatic variable selection (Zou & Hastie, 2005). It will also be unable to overcome the problem of multicollinearity in the given data. Hence, we take advantage of the property of the elastic net, which simultaneously achieves automatic variable selection, continuous shrinkage, and selection of the groups of correlated variables. It is like a "stretchable fishing net that retains all the big fish." We use the function `glmnet` and `cv.glmnet` in the package `glmnet` with a very high value of alpha (α) for variable selection and a very small value of lambda minimum for cross-validation of the model (J Friedman, Hastie, & Tibshirani 2010 & 2013). The function `cv.glmnet` runs `glmnet` nfolds (10) +1 times; the first to get the lambda sequence, and then the remainder to compute the fit with each of the folds omitted. The error is accumulated, and the average error and standard deviation over the folds is computed. Selection of the value for the lambda.min determines the minimum mean cross-validated error. Thereafter stepwise, multiple linear regression, based on Bayesian Information Criterion, was iteratively used to obtain only significant variables.

For a comparative study of these techniques, same parameters of seed= 5226, and `alpha`= 0.93 were selected, whereas the value of lambda.min was changed to calculate the optimum value of R-squared, adjusted R-squared and predicted R-squared.

First, we will use the complete data without removing any outliers from the original matrix of 144 observation, and 987 covariates to calculate the value of R-Squared, adjusted R-Squared and predicted R-Squared. These R-Squared values calculated with lambda.min (minimum mean cross-validated error) of 0.005 is given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.65	0.63	0.59	0.59	0.42	0.57
Adj.R.Squared	0.59	0.54	0.51	0.55	0.39	0.50
Pred.R.Squared	0.39	0.39	0.40	0.50	0.30	0.42

Second, since the original matrix of 144 observation and 987 covariates has 900 bad observations with value more than 100 and less than 0, we can replace them with the mean of the matrix of predictor variables mean. Then the values of R-Squared, adjusted R-Squared and predicted R-Squared using the same parameters and lambda.min = 0.0041 are as given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.79	0.57	0.67	0.56	0.70	0.56
Adj.R.Squared	0.74	0.50	0.61	0.51	0.65	0.49
Pred.R.Squared	0.63	0.40	0.55	0.44	0.54	0.38

To avoid the masking or swamping effects prevalent in linear regression models, we will use the robust linear regression to find a fit that is close to the fit we would have found without the outliers. Then we can identify the outliers by their significant deviation from that robust fit (Rousseeuw & Hubert, 2011). Instead of function step, in the package `stats`, we will use the function `lmrob` in the package `robustbase` to fit generalized linear models by robust methods. This function computes an MM-type regression estimator and the associated M-, S- and D estimators. M-estimation is an extension of the maximum likelihood estimate method and a robust estimation. S-estimation minimizes the scale of the residual from M-estimation (Susanti & Pratiwi, 2014). We have selected the setting as "KS2014," which uses the setting method = 'SMDM.' In this procedure, first estimate the regression parameter using S-estimation, followed by with M-estimation then proceed with a Design Adaptive Scale estimation and a final M-step (Maechler et al., 2016). The values of R-Squared, adjusted R-Squared, and predicted R-Squared using the same parameters and lambda.min = 0.009 are as given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.squared	0.15	0.15	0.16	0.46	0.29	0.25
Adj.R.Squared	0.13	0.13	0.15	0.44	0.27	0.21
Pred.R.Squared	0.10	0.08	0.12	0.38	0.16	0.14

Robust regression uses repeated median estimates to maintain up to 50% breakdown value. In other words even when nearly half of the data are outliers, robust regression can resist its effect. From a sample size of 144, there are 47, 38, 46, 41, 44 and 46 outliers of nitrogen,

potassium, phosphorus, magnesium, zinc and boron, respectively, with weights other than one, which is more than 30 percent of the data. Despite removing such a large number of outliers, it fails to remove all the wrong observations, because they lie within three standard deviations from the median. These explanations mentioned above might be the cause for such low values of R-Squared, adjusted R-Squared and predicted R-Squared.

8.3 Multicollinearity

The spectral reflectance is measured at leaf or canopy level over the wavelength from 330 to 2510 nanometers, and the nutrient analysis was conducted at the petiole level in the grapevine dataset. Since it is a case of multiple linear regression of the same type of data measured at a close interval of 1.5 to 2.7 nanometers, hence we can expect the predictor variables to be highly correlated. Data visualization using a correlation matrix plot can help to gain a better understanding of the problem of collinearity in the grapevine dataset. These pairwise correlations can give an idea of which attributes change together. The areas covered by the sector shows the absolute value of corresponding correlation coefficients. The bigger the sector, the larger the correlation. The diagonal of the matrix plot are perfectly positively correlated because it illustrates the correlation of each attribute with itself. Also, the area of each pie is shaded blue or red depending on the sign of the correlation, and with the strength of color scaled 0–100% in proportion to the magnitude of the correlation. Blue represents positive correlation and red negative.

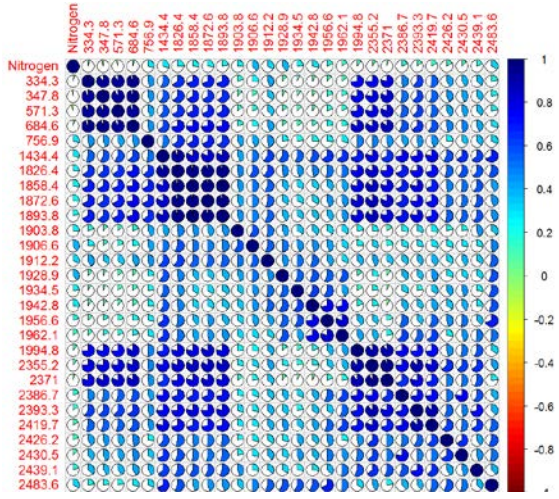


Figure 8.3: Correlation plot of Wavelength for Nitrogen

Figure 8.3 to Figure 8.8 shows the correlation matrix plot for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The correlation matrices with the response variables of the grapevine dataset are symmetrical and perfectly positively correlated along the diagonal. The range of pairwise correlation in percentage is given in Table 8.1.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Max Correlation	98%	98%	98%	94%	97%	98%
Min Correlation	-5%	2%	-32%	-30%	-9%	-24%

Table 8.1: Max and Min correlation with the response variables of the grapevine dataset

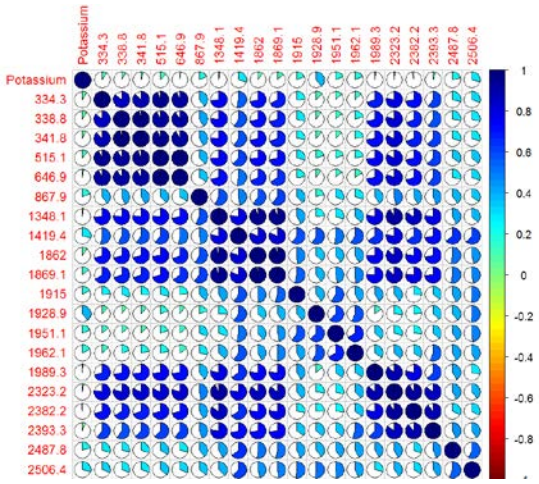


Figure 8.4: Correlation plot of Wavelength for Potassium

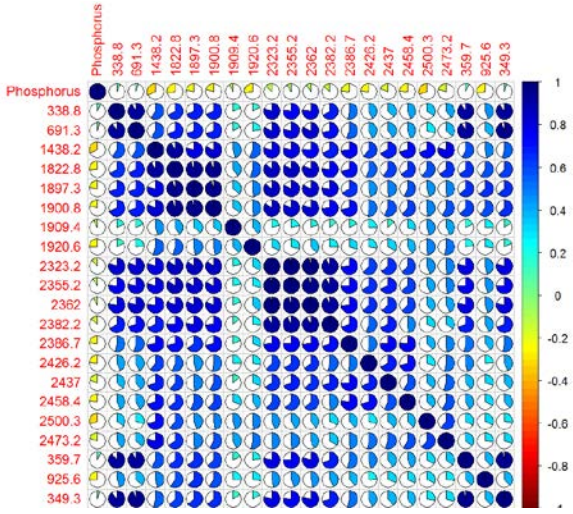
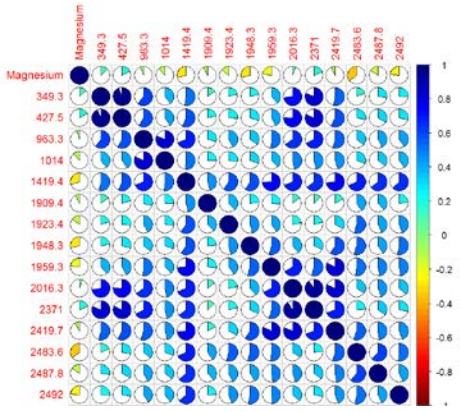
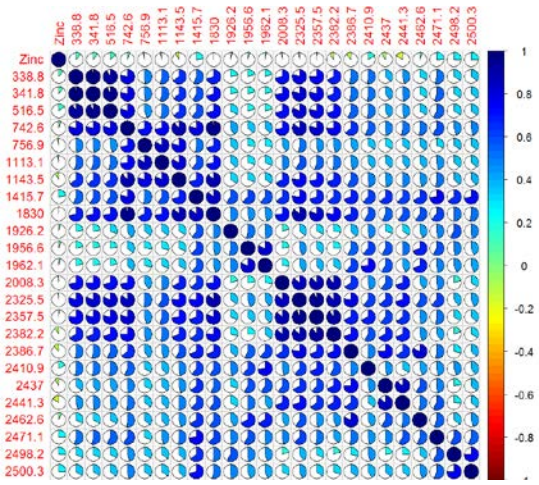


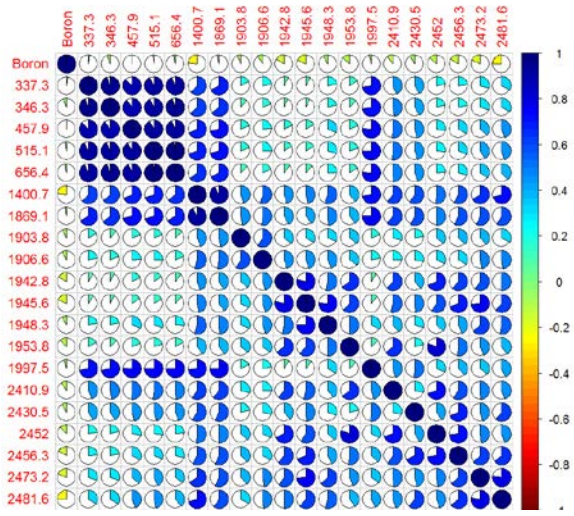
Figure 8.5: Correlation plot of Wavelength for Phosphorus



`Figure 8.6:` Correlation plot of Wavelength for Magnesium



`Figure 8.7:` Correlation plot of Wavelength for Zinc



`Figure 8.8`: Correlation plot of Wavelength for Boron

Since the predictors are highly correlated, we need to calculate the variance inflation factor (VIF), which quantifies the severity of multicollinearity in a multiple linear regression. The square root of VIF gives the magnitude of standard error as compared with uncorrelated predictors. Hence, lower levels of VIF is desirable, as higher levels of VIF adversely affect the results associated with a multiple regression analysis. However, for the grapevine dataset, to achieve the VIF of less than 10, the model removes all the variables which are highly correlated. This result in a very low value of R-Squared adjusted R-Squared and predicted R-Squared for all the six nutrients as given below.

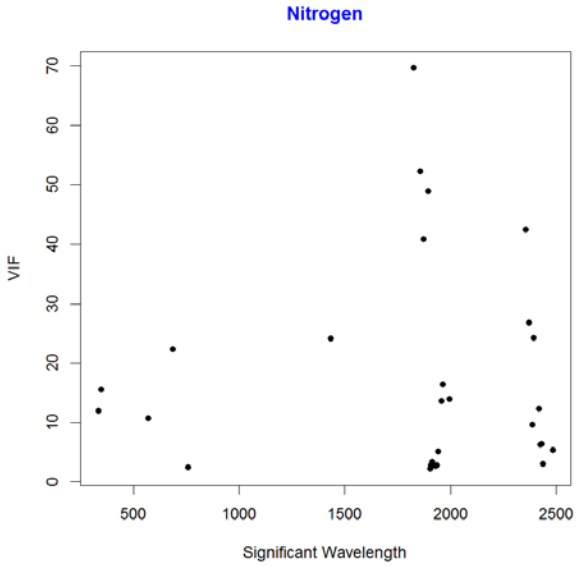
	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.26	0.21	0.35	0.50	0.56	0.29
Adj.R.Squared	0.24	0.19	0.31	0.45	0.51	0.28
Pred.R.Squared	0.20	0.16	0.22	0.40	0.39	0.27

Elastic net is known to select groups of correlated variables, which does not affect the predictability of the model. Since multicollinearity does not influence the overall fit of the model or produce wrong predictions, hence, the upper limit of the VIF has been selected as 80. By selecting the upper limit of VIF as 80, elastic net selects a few the predictors with very high, but the median VIF is around 10. The process mentioned above ensures the optimum value of R-Squared adjusted R-Squared and predicted R-Squared for all the six nutrients as given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.79	0.57	0.67	0.56	0.7	0.56
Adj.R.Squared	0.74	0.5	0.61	0.51	0.65	0.49
Pred.R.Squared	0.63	0.4	0.55	0.44	0.54	0.38

Significant Wavelength (nm): 334.3, 347.8, 571.3, 684.6, 756.9, 1434.4, 1826.4, 1858.4, 1872.6, 1893.8, 1903.8, 1906.6, 1912.2, 1928.9, 1934.5, 1942.8, 1956.6, 1962.1, 1994.8, 2355.2, 2371, 2386.7, 2393.3, 2419.7, 2426.2, 2430.5, 2439.1, 2483.6

Variance Inflation Factor (VIF): 11.98, 15.5, 10.8, 22.31, 2.49, 24.2, 69.73, 52.29, 40.84, 48.98, 2.21, 2.83, 3.34, 2.74, 2.79, 5.15, 13.61, 16.48, 14.02, 42.45, 26.78, 9.65, 24.31, 12.31, 6.3, 6.44, 3.03, 5.33



`Figure 8.9:` Scatterplot of VIF against Wavelength for Nitrogen

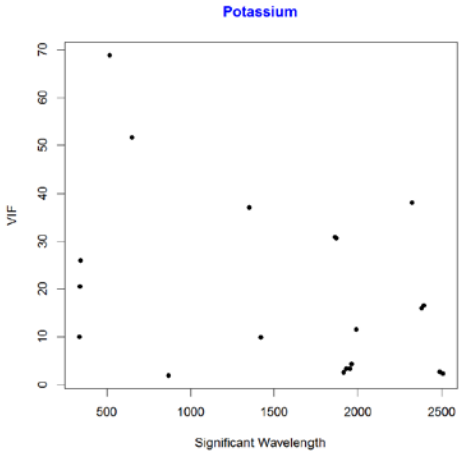
Significant Wavelength (nm): 334.3, 338.8, 341.8, 515.1, 646.9, 867.9, 1348.1, 1419.4, 1862, 1869.1, 1915, 1928.9, 1951.1, 1962.1, 1989.3, 2323.2, 2382.2, 2393.3, 2487.8, 2506.4

Variance Inflation Factor (VIF): 9.96, 20.58, 26.02, 68.94, 51.73, 1.85, 37.03, 9.87, 30.85, 30.67, 2.58, 3.32, 3.37, 4.29, 11.53, 38.08, 15.97, 16.54, 2.71, 2.36

`Figure 8.9` to `Figure 8.14` displays the scatterplot of VIF against the wavelengths for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The concentration of significant wavelengths for the response variables for the grapevine dataset can be seen to have VIF around 10. Certain significant wavelengths have very high VIF; however, the median VIF of significant predictors (wavelength) has been tabulated in `table 8.2`.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Median of VIF	Around 10	Around 14	Around 10	Around 4	Around 8	Around 10

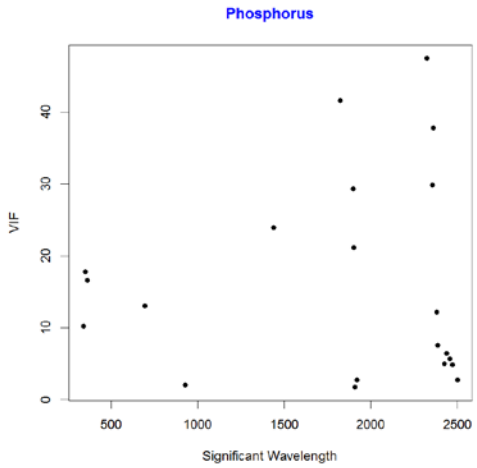
`Table 8.2:` Median VIF of significant predictors for response variable of grapevine dataset



`Figure 8.10`: Scatterplot of VIF against Wavelength for Potassium

Significant Wavelength (nm): 338.8, 691.3, 1438.2, 1822.8, 1897.3, 1900.8, 1909.4, 1920.6, 23 23.2, 2355.2, 2362, 2382.2, 2386.7, 2426.2, 2437, 2458.4, 2500.3, 2473.2, 359.7, 925.6, 349.3

Variance Inflation Factor (VIF): 10.21, 13.06, 23.98, 41.64, 29.33, 21.15, 1.7, 2.72, 47.54, 29.85, 37.85, 12.19, 7.55, 5.03, 6.41, 5.69, 2.74, 4.86, 16.56, 2.01, 17.82



`Figure 8.11:` Scatterplot of VIF against Wavelength for Phosphorus

Significant Wavelength (nm): 349.3, 427.5, 963.3, 1014, 1419.4, 1909.4, 1923.4, 1948.3, 1959.3, 2016.3, 2371, 2419.7, 2483.6, 2487.8, 2492

Variance Inflation Factor (VIF): 10.63, 10.88, 6.49, 3.87, 7.42, 1.54, 2.04, 2.12, 4.05, 10.19, 9, 7. 31, 2.93, 2.01, 2.07

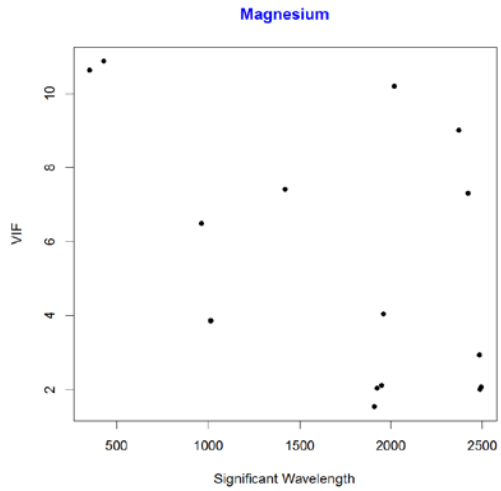


Figure 8.12: Scatterplot of VIF against Wavelength for Magnesium

Significant Wavelength (nm): 338.8, 341.8, 516.5, 742.6, 756.9, 1113.1, 1143.5, 1415.7, 1830, 1926.2, 1956.6, 1962.1, 2008.3, 2325.5, 2357.5, 2382.2, 2386.7, 2410.9, 2437, 2441.3, 2462.6, 2471.1, 2498.2, 2500.3

Variance Inflation Factor (VIF): 22.2, 27.55, 13.96, 35.85, 5.56, 4.83, 8.57, 10.78, 37.89, 2.62, 6.03, 7.93, 9.11, 36.37, 28.53, 13.12, 10.33, 8.68, 8.51, 8.55, 6.25, 4.21, 3.9, 3.37

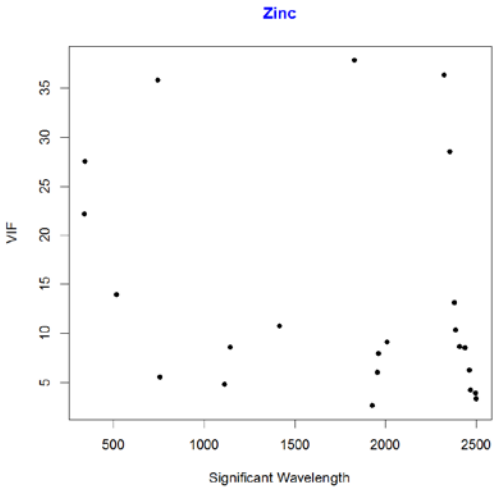


Figure 8.13: Scatterplot of VIF against Wavelength for Zinc

Significant Wavelength (nm): 337.3, 346.3, 457.9, 515.1, 656.4, 1400.7, 1869.1, 1903.8, 1906.6, 1942.8, 1945.6, 1948.3, 1953.8, 1997.5, 2410.9, 2430.5, 2452, 2456.3, 2473.2, 2481.6

Variance Inflation Factor (VIF): 11.82, 15.91, 10.24, 46.85, 57.18, 11.59, 10, 1.91, 2.34, 4.22, 13.99, 6.83, 6.61, 6.77, 6.61, 5.97, 9, 6.54, 4.81, 4.78

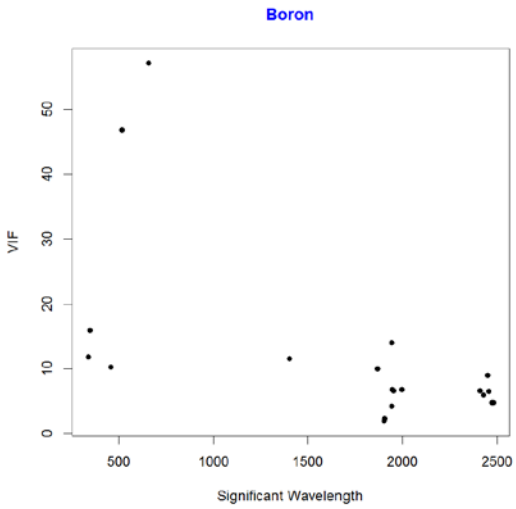
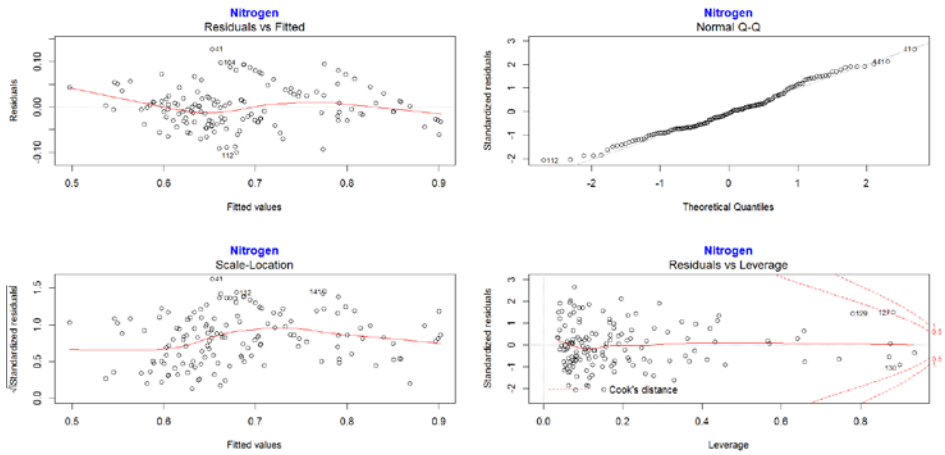


Figure 8.14: Scatterplot of VIF against Wavelength for Boron

The elastic net is known to select highly correlated predictors due to grouping effect, without compromising its predictive ability. Hence, we can notice high VIF due to the very high pair wise correlation between some predictor variables. Therefore, such high VIF may be acceptable in this case.

8.4 Residual Analysis

So far, we have checked regression results, such as slope coefficients, p-values, multicollinearity and R-Squared to understand fitment of a model for the given data. Residual analysis is a useful class of techniques for the evaluation of the goodness of fit. Residuals are leftover, after fitting a model (predictors) to data, and they could reveal unexplained patterns in the data. Examining the underlying assumptions is important since most linear regression estimators require a correctly specified regression function and independent and identically distributed residual to be consistent. A residual plot is a nice way to show the residuals on the vertical axis and the independent variable on the horizontal axis. When the points in a residual plot are randomly dispersed around the horizontal axis, then the linear regression model is considered appropriate for the dataset.

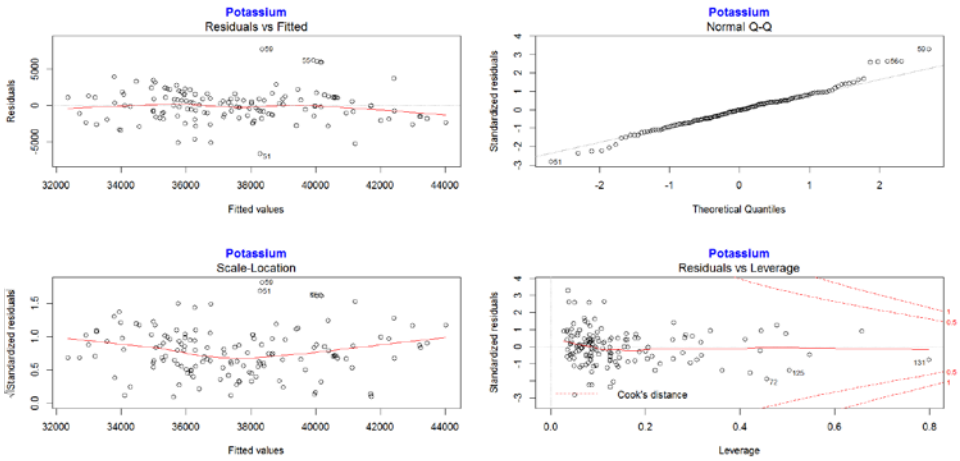


`Figure 8.15:` Residual Plot of Nitrogen

For the residual vs. fitted plot for nitrogen, potassium, phosphorus, magnesium, zinc and boron are shown in the `Figure 8.15` to `Figure 8.20` in sequence. The observations number with large values of standardized residual, which may be considered as outliers, shown in `Table 8.3`.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Outliers	41, 112 and 141	51, 55 and 59	38, 108 and 110	37, 38 and 40	44, 68 and 70	51, 54 and 56

`Table 8.3:` Outliers for response variable of grapevine dataset



`Figure 8.16:` Residual Plot of Potassium

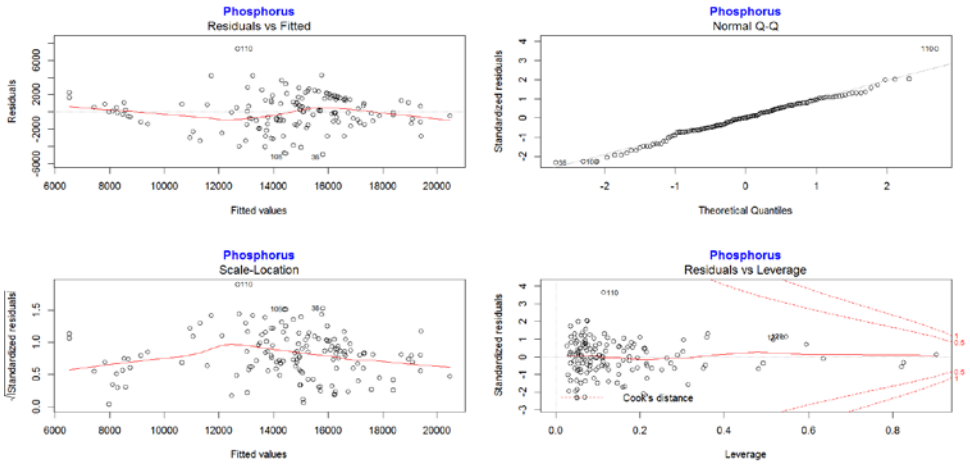


Figure 8.17: Residual Plot of Phosphorus

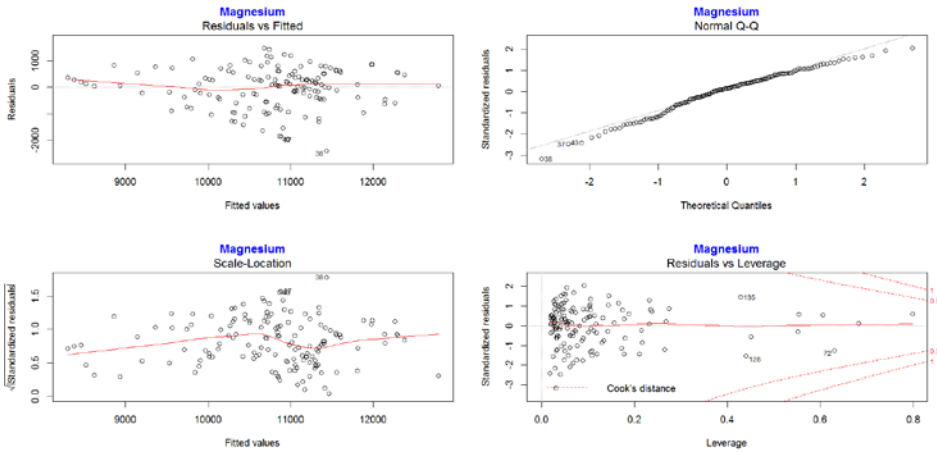


Figure 8.18: Residual Plot of Magnesium

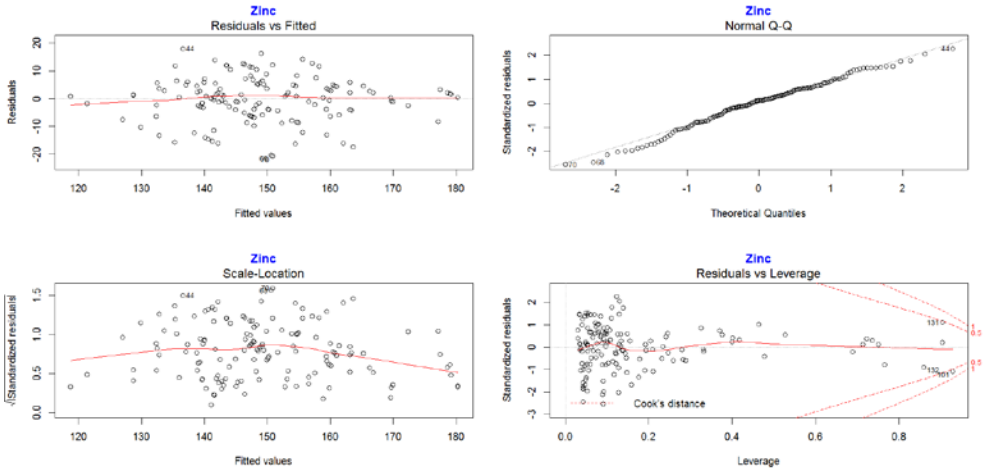


Figure 8.19: Residual Plot of Zinc

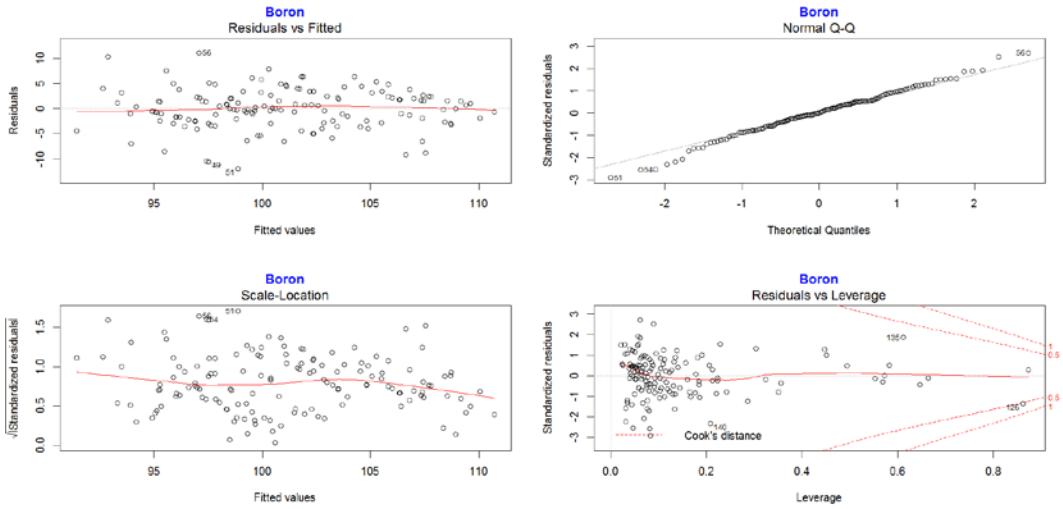


Figure 8.20: Residual Plot of Boron

However, their distribution appears to be equally spread around, a horizontal line without any distinctly discernable patterns, which indicates that we do not have non-linear relationships. The normal Q-Q plot in the Figure 8.15 to Figure 8.20 shows that residuals are normally distributed without much deviation, as desired. The scale-Location plot indicates that the residuals are more or less equally (randomly) spread along the ranges of predictors. Almost all the residuals are within two standard deviations. The explanation mentioned above satisfies the assumption of equal variance (homoscedasticity). The residuals vs. leverage plot do indicate residual for the certain observations as given in Table 8.4 have large Cook's distance and possible influential cases, but their exclusion would not alter the regression results.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Influential cases	127, 129 130	72, 125 and 131	110, 124 and 129	72, 128 and 135	101, 131 and 132	126, 135 and 140

Table 8.4: Influential cases for response variable of grapevine dataset

Chapter 9

Variable Selection of

Riesling Bloom Leaf Analysis

9.1 Introduction

In chapter 8, we have seen that grapevine dataset, about the petiole chemical analysis of the Riesling variety, as modeled via the individual leaf reflectance during the bloom period, has 900 bad observations. Complete scrutiny of this grapevine dataset has revealed 828 and 72 observations with values more than 100 and less than zero respectively. Due to the presence of certain extreme outliers, the standard deviation is high, resulting in a high Coefficient of Variation of around 125 percent. Hence, we replace all values less than 0 and more than 100 percent by the mean value of the matrix, even though it is within one and two standard deviations, respectively. We have seen that the best value of R-squared adjusted R-squared and predicted R-squared were obtained by accepting the slightly higher value of VIF.

9.2 Methods for Wavelength Selection

With many predictors, fitting the full model without penalization will result in large prediction intervals, and the least square regression estimator may not uniquely exist. The coefficients for some predictors may not be significantly different from 0, and hence they may not influence the prediction of the response variable. A proper choice of selection methods and under appropriate conditions will help to build consistent models to select variables and estimate coefficients simultaneously, avoid model overfitting, and obtain satisfactory prediction accuracy. Since the number of predictors is more than the sample size, hence to provide a sparser representation of the data and a reasonable statistical model, we explore four efficient algorithms for variable selection. In this thesis, we consider the feature selection under optimization algorithms for penalized regression methods and functional regression.

9.3 Penalized (Pseudo-) Likelihood Approach (Elastic Net) using package glmnet

First, we take advantage of algorithms for estimation of linear models with the convex penalized (pseudo-) likelihood approach. The models include elastic net for high-dimensional correlated variables, which uses a mixture of the ℓ_1 (lasso) and ℓ_2 (ridge regression) penalties to achieve a sparse solution. The regularization path is computed for the elastic net penalty at a grid of values for the regularization parameter lambda. It has the effect of averaging wavelengths that are highly correlated and then entering the averaged wavelengths into the model. The algorithm is used to compute of the entire path of solutions for each method, at 100 values of the regularization parameter spaced on the log-scale.

These algorithms use cyclical coordinate descent, computed along a regularization path developed in the package glmnet (J Friedman et al., 2013 & 2010). The regularization path is computed for the elastic net penalty at a grid of values for the regularization parameter lambda. For "Gaussian," (this case) glmnet standardizes y to have unit variance before computing its lambda sequence and then removes standardization to yield the resulting coefficients. The coefficients for any predictor variables with zero variance are set to zero for all values of lambda. The algorithm used for this loops through the number of observations every time an inner product is computed. Coordinate descent fits the elastic net sequence of models implied by lambda. The function cv.glmnet has been used for cross-validation and the lambda.min for obtaining the value of λ that gives the minimum mean cross-validated error. The default value of lambda.min.ratio, which is the smallest value for lambda, as a fraction of lambda.max (the data derived entry value) is 0.01 when the sample size is less than some variables. A small value of lambda.min.ratio will lead to a full model in this case. Hence, the selection of the value of lambda.min.ratio was done considering these factors, and to obtain the best possible value of R–Squared, adjusted R–Squared and predicted R–Squared, the value of alpha and lambda minimum ratio were varied. For fair comparison, seed of 5226, alpha = 0.93, lambda.min = 0.0041 and lambda.min.ratio = 0.0041 has been considered after replacing all the outliers with the mean of the input matrix in the data. Each curve represents a coefficient in the regression model. The x-axis is a function of lambda, the regularization penalty parameter. The y-axis gives the value of the coefficient.

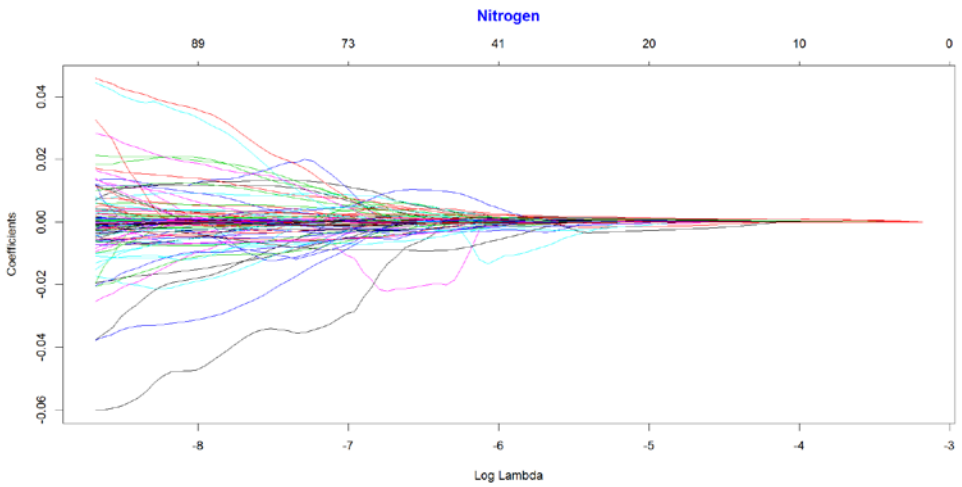


Figure 9.1: Model Coefficient Path using Elastic Net for Nitrogen

The Figure 9.1, Figure 9.4, Figure 9.7, Figure 9.10, Figure 9.13 and Figure 9.16 displays Model Coefficient Path using Elastic Net for nitrogen, potassium, phosphorus, magnesium, zinc and boron respectively. These figures demonstrate, how the coefficients of the nutrients enter the model (become non-zero) as lambda changes. Most of the variables have coefficients close to zero, which indicates high collinearity. However, the elastic net is capable of handling such multicollinearity, by the grouping effect.

The red dots are the mean computed using leave-one-out cross-validation. Confidence intervals represent error estimates for the loss metric (red dots). The vertical lines show the locations of λ_{min} and λ_{1se}. The numbers across the top are the number of nonzero coefficient estimates. The best λ to use is either the one giving minimum MSE or the largest λ, such that error is within one standard deviation from the minimum. However, in this thesis, the value of λ_{min} has been used to calculate the non-zero variables.

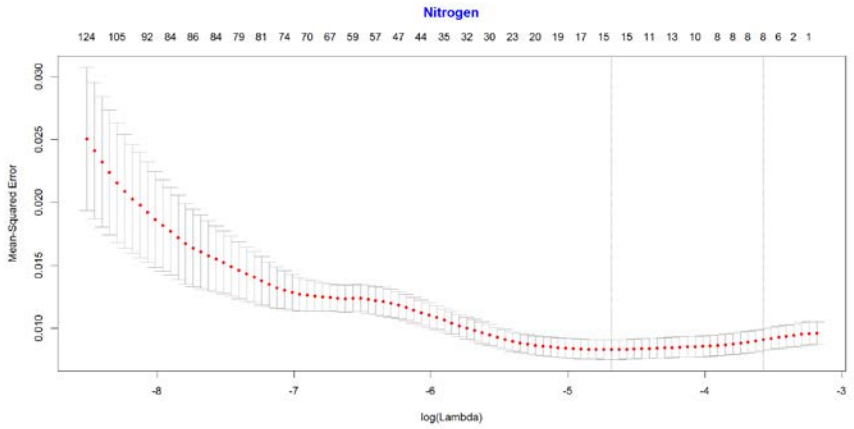


Figure 9.2: Mean-Squared Error and log (λ) using Elastic Net for Nitrogen

The Figure 9.2, Figure 9.5, Figure 9.8, Figure 9.11, Figure 9.14 and Figure 9.17 displays the plot of Mean-Squared Error against the log of lambda for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The sharp drop in mean square error around log lambda minimum and log lambda 1se explains a substantial fraction of the variability in all the six nutrients. We can also notice that the standard errors are initially wide, and then it narrows down. However, for the further study, we will pursue with lambda min. The value of lambda minimum, log of lambda minimum, lambda displaced by one standard error (SE), a log of lambda 1 SE corresponding to the minimum value of Mean Square Error for the six nutrients obtained from these figures have been tabulated in Table 9.1. Some variables with nonzero coefficients corresponding to the value of the log of lambda minimum and a log of lambda 1 SE has been included in this table.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Lambda minimum	0.01	284.07	283.56	49.87	0.75	0.31
Log Lambda minimum	-4.68	5.65	5.65	3.91	-0.29	-1.16
No. nonzero coefficients (λ_{min})	15	14	18	25	25	14
Lambda 1SE	0.03	1344.62	1074.87	143.23	1.71	0.58
Log Lambda 1SE	-3.57	7.20	6.98	4.96	0.54	-0.55
No. nonzero coefficients (λ_{1SE})	8	1	6	12	15	10

Table 9.1: Lambda values corresponding to the minimum MSE

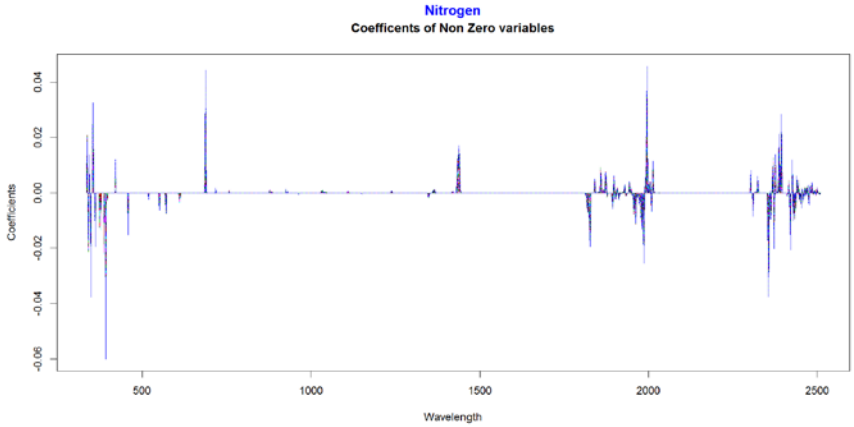


Figure 9.3: Coefficients of Non-Zero Variables for Nitrogen

Figure 9.3 shows the regression coefficients of 986 variables obtained by the elastic net. Based on the value of lambda min of 0.01, there are 77 non-zero coefficients are selected into regression model for the prediction of nitrogen; remaining coefficients have shrunk to be zero. The grouping or clustering of the wavelengths is clearly visible. Grouping of the wavelengths into five clusters and one lone variable are clearly visible. Cluster 1: 334.3, 337.3, 338.8, 340.3, 347.8, 373.1, 386.4 and 392.3 nm. Cluster 2: 568.5, 569.9, 571.3, 684.6, 685.9, 687.3, 756.9, 877.5, 882.2, and 884.6 nm. Lone variable (wavelength) is 1109.2 nm. Cluster 3: 1419.4, 1434.4 and 1438.2 nm. Cluster 4: 1819.3, 1822.8, 1826.4, 1854.9, 1858.4, 1865.5, 1872.6, 1893.8, 1903.8, 1906.6, 1912.2, 1920.6, 1926.2, 1928.9, 1931.7, 1934.5, 1942.8, 1948.3, 1956.6, 1959.3, 1962.1, 1978.5, 1994.8, 1997.5 and 2016.3 nm. Cluster 5: 2355.2, 2368.8, 2371, 2386.7, 2388.9, 2393.3, 2410.9, 2415.3, 2419.7, 2426.2, 2430.5, 2432.7, 2437, 2439.1, 2443.4, 2447.7, 2452, 2462.6, 2464.8, 2473.2, 2475.3, 2483.6, 2485.7, 2487.8, 2492, 2496.1, 2498.2, 2500.3, 2504.4 and 2506.4 nm.

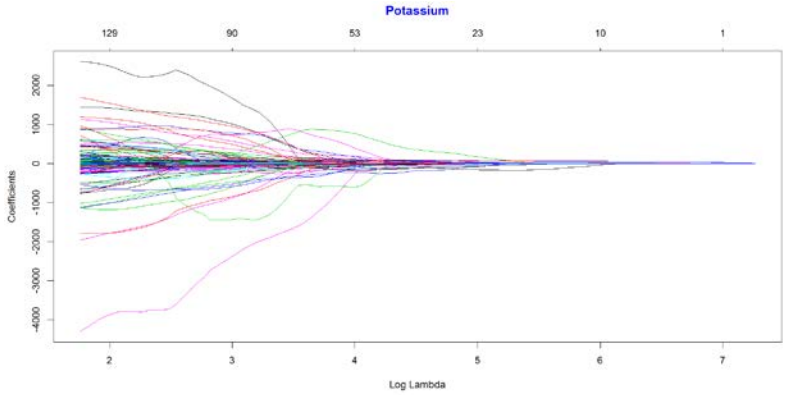


Figure 9.4: Model Coefficient Path using Elastic Net for Potassium

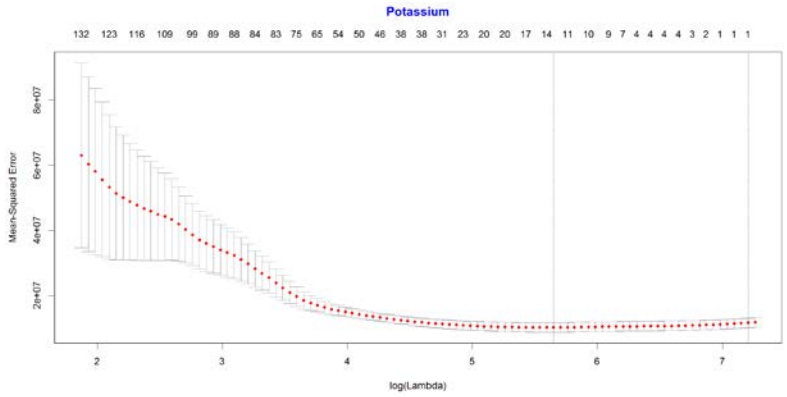


Figure 9.5: Mean-Squared Error and log (λ) using Elastic Net for Potassium

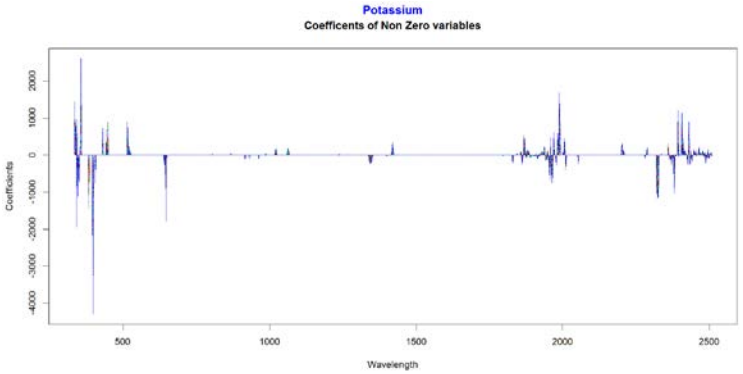


Figure 9.6: Coefficients of Non-Zero Variables for Potassium

The Figure 9.6 shows the regression coefficients of 986 variables obtained by elastic net. Based on the value of lambda min of 284, there are 82 non-zero coefficients are selected into regression model for the prediction of potassium; remaining coefficients have shrunk to be zero. The grouping or clustering of the wavelengths is clearly visible. Grouping of the wavelengths into five clusters are clearly visible. Cluster 1**:** 334.3, 337.3, 338.8, 340.3, 341.8, 356.8, 383.5, 398.2, 443.5, 449.3, 515.1, 516.5, 517.9, 638.7, 640.1, 641.5 and 646.9 nm. Cluster 2: 855.8, 867.9, 914, 926.8, 963.3, 987.3, and 1063.5 nm. Cluster 3: 1344.3, 1348.1 and 1419.4 nm. Cluster 4: 1826.4, 1830, 1858.4, 1862, 1869.1, 1876.1, 1890.2, 1897.3, 1903.8, 1912.2, 1915, 1928.9, 1934.5, 1937.3, 1940, 1942.8, 1945.6, 1948.3, 1951.1, 1956.6, 1962.1, 1989.3 and 2005 nm. Cluster 5: 2323.2, 2325.5, 2359.8, 2371, 2373.3, 2380, 2382.2, 2393.3, 2406.6, 2410.9, 2413.1, 2419.7, 2441.3, 2449.9, 2458.4, 2462.6, 2469, 2471.1, 2477.4, 2479.5, 2483.6, 2485.7, 2487.8, 2492, 2494.1, 2496.1, 2498.2, 2500.3, 2502.3, 2504.4, 2506.4 and 2508.5 nm.

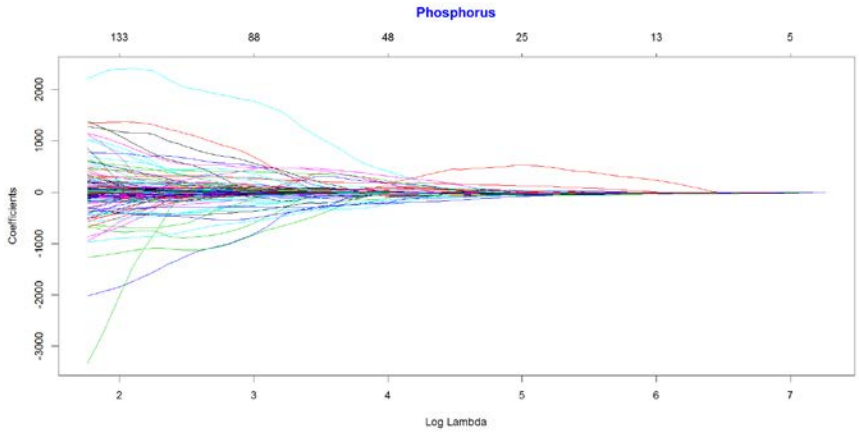


Figure 9.7: Model Coefficient Path using Elastic Net for Phosphorus

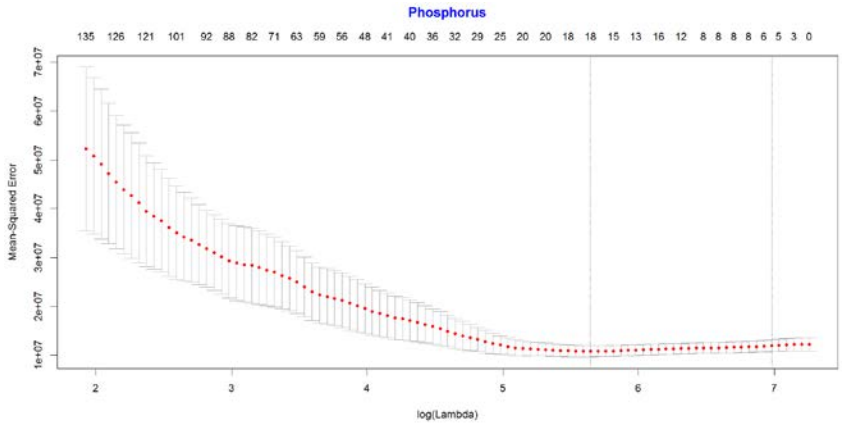


Figure 9.8: Mean-Squared Error and log (λ) using Elastic Net for Phosphorus

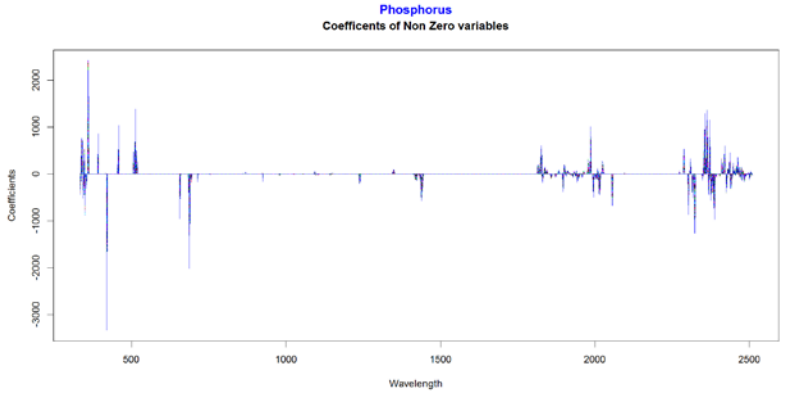


Figure 9.9: Coefficients of Non-Zero Variables for Phosphorus

Figure 9.9 shows the regression coefficients of 986 variables obtained by the elastic net. Based on the value of lambda min of 284, there are 71 non-zero coefficients are selected into regression model for the prediction of phosphorus; remaining coefficients have shrunk to be zero.

The grouping or clustering of the wavelengths is clearing visible. Grouping of the wavelengths into six clusters are clearly visible. Cluster 1: 334.3, 338.8, 340.3, 347.8, 349.3 and 359.7 nm. Cluster 2: 687.3, 688.6, 691.3 and 692.6 nm. Cluster 3: 855.8, 860.7, 925.6, 967.9 and 979.1 nm. Cluster 4: 1423.2, 1434.4, 1438.2 and 1441.9 nm. Cluster 5: 1822.8, 1858.4, 1893.8, 1897.3, 1900.8, 1903.8, 1906.6, 1909.4, 1912.2, 1915, 1920.6, 1928.9, 1931.7, 1934.5, 1942.8, 1948.3, 1951.1, 1978.5, 1986.6, 2016.3 nm. Cluster 6: 2323.2, 2353, 2355.2, 2362, 2368.8, 2371, 2382.2, 2386.7, 2404.4, 2410.9, 2413.1, 2426.2, 2430.5, 2437, 2439.1, 2441.3, 2447.7, 2452, 2458.4, 2462.6, 2464.8, 2473.2, 2479.5, 2483.6, 2485.7, 2489.9, 2492, 2498.2, 2500.3, 2502.3, 2506.4 and 2508.5 nm.

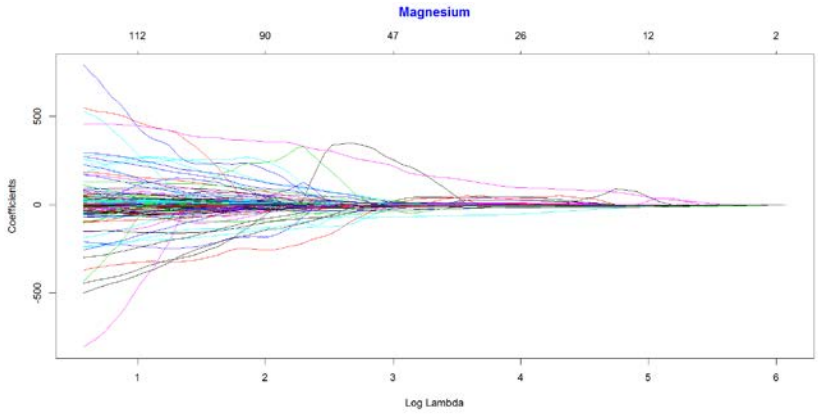


Figure 9.10: Model Coefficient Path using Elastic Net for Magnesium

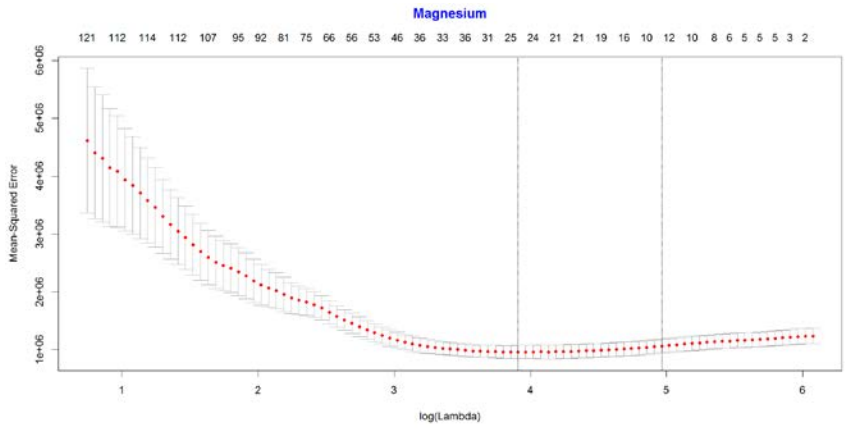


Figure 9.11: Mean-Squared Error and log (λ) using Elastic Net for Magnesium

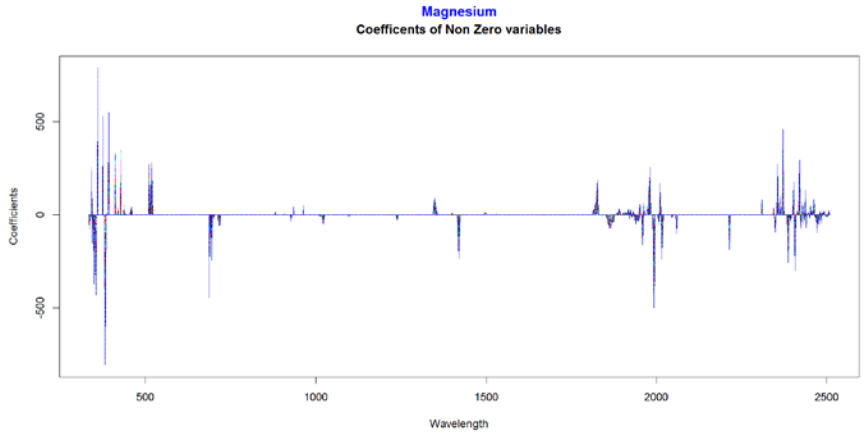


Figure 9.12: Coefficients of Non-Zero Variables for Magnesium

The Figure 9.12 shows the regression coefficients of 986 variables obtained by the elastic net. Based on the value of lambda min of 49.9, there are 58 non-zero coefficients are selected into regression model for the prediction of magnesium; remaining coefficients have shrunk to be zero. The grouping or clustering of the wavelengths is clearly visible. Grouping of the wavelengths into six clusters and one lone variable are clearly visible. Cluster 1: 337.3, 340.3, 341.8, 343.3, 349.3, 411.4 and 427.5 nm. Cluster 2: 695.3, 699.2, 700.6, 712.5, 717.8 and 719.1 nm. Cluster 3: 963.3, 1010.2, 1014, 1017.8, 1021.6, 1097.8 and 1419.4 nm. Cluster 4: 1858.4, 1862, 1903.8, 1909.4, 1912.2, 1917.8, 1920.6, 1923.4, 1928.9, 1931.7, 1937.3, 1948.3, 1959.3, 1962.1, 2010.9 and 2016 nm. Cluster 5: 2343.9, 2355.2, 2371, 2373.3, 2384.4, 2419.7, 2432.7, 2437, 2439.1, 2466.9, 2469, 2471.1, 2481.6, 2483.6, 2487.8, 2492, 2494.1, 2500.3, 2502.3, 2504.4, 2506.4 and 2508.5 nm.

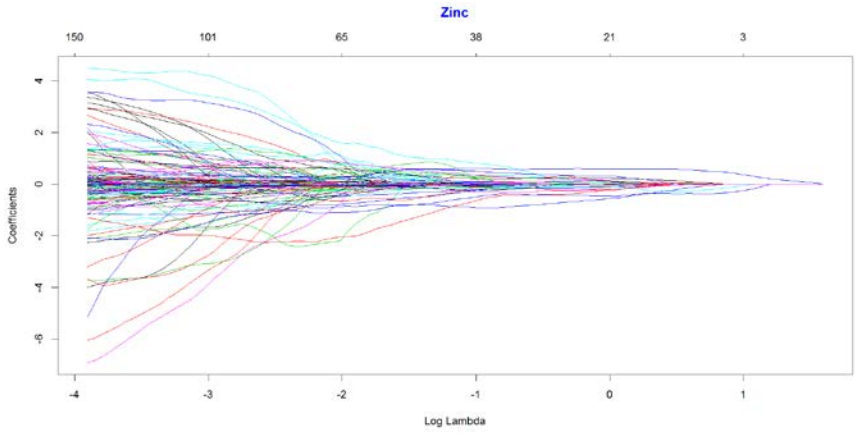


Figure 9.13: Model Coefficient Path using Elastic Net for Zinc

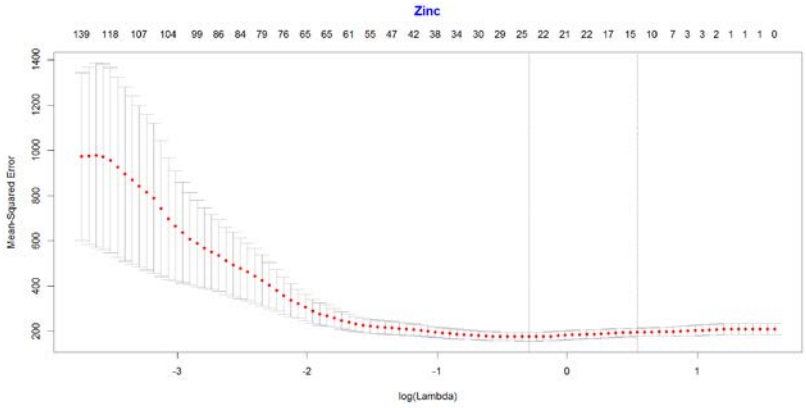


Figure 9.14: Mean-Squared Error and log (λ) using Elastic Net for Zinc

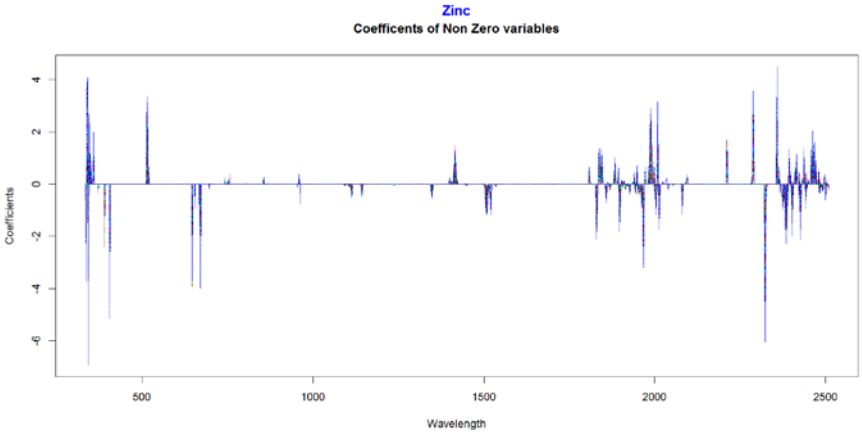


Figure 9.15: Coefficients of Non-Zero Variables for Zinc

The Figure 9.15 shows the regression coefficients of 986 variables obtained by elastic net. Based on the value of lambda min of 0.75, there are 65 non-zero coefficients are selected into regression model for the prediction of boron; remaining coefficients have shrunk to be zero. The grouping or clustering of the wavelengths is clearly visible. Grouping of the wavelengths into six clusters are clearly visible. Cluster 1: 334.3, 338.8, 340.3, 341.8, 390.9 nm. Cluster 2: 516.5, 517.9, 742.6, 756.9 nm. Cluster 3: 959.9, 1094, 1113.1, 1143.5 nm. Cluster 4: 1348.1, 1411.9, 1415.7, 1505.3, 1509.1 nm. Cluster 5: 1830, 1897.3, 1903.8, 1906.6, 1909.4, 1912.2, 1915, 1920.6, 1923.4, 1926.2, 1928.9, 1934.5, 1940, 1942.8, 1948.3, 1951.1, 1956.6, 1962.1, 1983.9, 1992.1, 2008.3 nm. Cluster 6: 2323.2, 2325.5, 2357.5, 2377.7, 2382.2, 2386.7, 2402.1, 2410.9, 2421.9, 2426.2, 2437, 2441.3, 2445.6, 2458.4, 2462.6, 2471.1, 2479.5, 2487.8, 2489.9, 2494.1, 2496.1, 2498.2, 2500.3, 2502.3, 2504.4, 2508.5 nm.

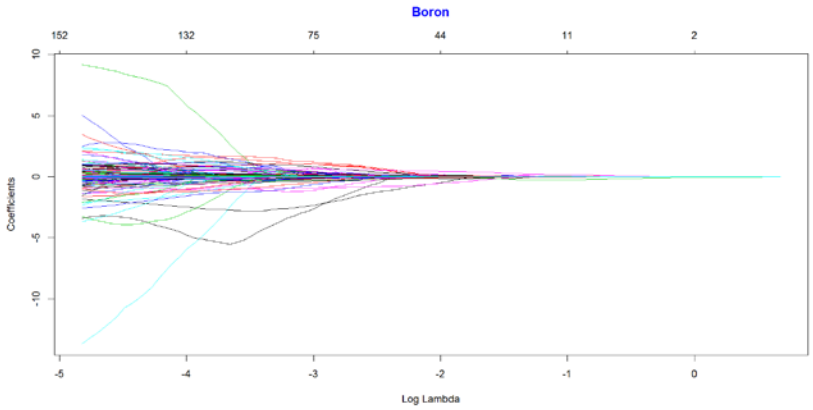


Figure 9.16: Model Coefficient Path using Elastic Net for Boron

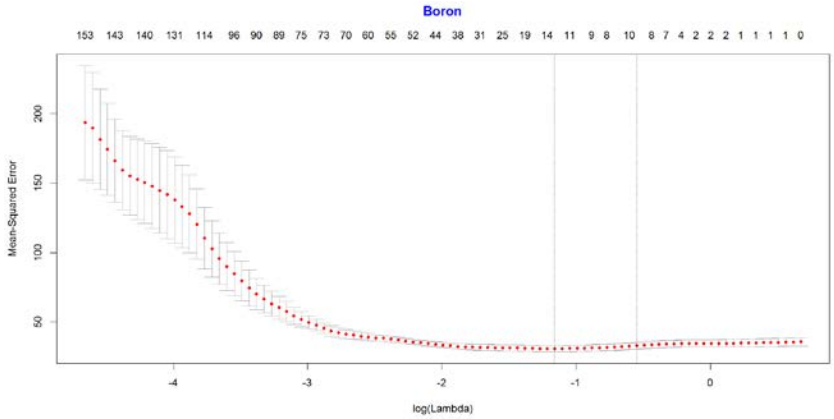


Figure 9.17: Mean-Squared Error and log (λ) using Elastic Net for Boron

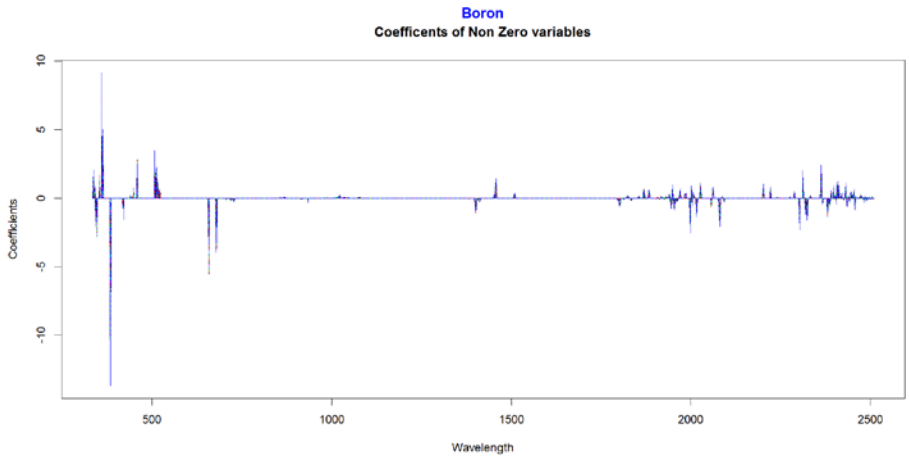


Figure 9.18: Coefficients of Non-Zero Variables for Boron

The Figure 9.18 shows the regression coefficients of 986 variables obtained by Elastic net. Based on the value of lambda min of 0.31, there are 81 non-zero coefficients are selected into regression model for the prediction of boron; remaining coefficients have shrunk to be zero. The grouping or clustering of the wavelengths is clearly visible. Grouping of the wavelengths into six clusters and a lone variable are clearly visible. Cluster 1: 335.8, 337.3, 340.3, 346.3, 353.8 nm. Cluster 2: 449.3, 457.9, 512.2, 513.7, 515.1, 516.5, 517.9, 519.3 nm. Cluster 2: 656.4, 657.7, 705.9, 707.2, 708.5, 709.8, 713.8, 715.1, 716.5, 717.8 nm. Cluster 3: 855.8, 860.7, 867.9, 870.3, 914, 933.7 nm. A lone wavelength is 1400.7 nm. Cluster 4: 1854.9, 1869.1, 1903.8, 1906.6, 1909.4, 1915, 1917.8, 1920.6, 1928.9, 1931.7, 1934.5, 1940, 1942.8, 1945.6, 1948.3, 1951.1, 1953.8, 1956.6, 1997.5 nm. Cluster 5: 2359.8, 2380, 2386.7, 2406.6, 2410.9, 2413.1, 2417.5, 2428.4, 2430.5, 2432.7, 2437, 2441.3, 2449.9, 2452, 2454.1, 2456.3, 2471.1, 2473.2, 2475.3, 2477.4, 2481.6, 2485.7, 2487.8, 2489.9, 2492, 2496.1, 2498.2, 2500.3, 2502.3, 2504.4, 2506.4, 2508.5 nm.

Thereafter, stepwise multiple linear regression, based on Bayesian Information Criterion, was iteratively used to obtain only significant variables with maximum variation inflation factor (VIF) of 100. The process mentioned above has ensured that most of the variables are around VIF of 10. The value of R–Squared, adjusted R–Squared and predicted R–Squared for the six nutrients are given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.79	0.57	0.67	0.56	0.70	0.56
Adj.R.Squared	0.74	0.50	0.61	0.51	0.65	0.49
Pred.R.Squared	0.63	0.40	0.55	0.44	0.54	0.38

9.4 Minimax Concave Penalty using package ncvreg

Second, we used group descent algorithms for nonconvex penalized linear regression models for high-dimensional regression and variable selection. To improve the efficiency of algorithms and to achieve simultaneous selection consistency and asymptotic unbiasedness, we fit the minimax concave penalty (MCP) in the package ncvreg (Breheny & Breheny, 2016). Estimation using MCP models depends on the choice of the tuning parameters gamma (γ) and lambda (λ). The value of lambda is usually obtained using cross-validation. However, cross-validation is computationally intensive, particularly when performing over a two-dimensional grid of values for γ and λ, some of which may not possess convex objective functions. The value of γ

is selected so that it produces parsimonious models while circumventing the pitfalls mentioned above for non-convexity (Breheny & Huang, 2011).

In linear regression, the scaling factor by which solutions are modified toward their unpenalized solution is a constant [$1 - 1/\gamma$ for MCP] for all values of λ and for each covariate. Since for global convexity: γ must be greater than $1/c_*$ for MCP, where c_* denotes the minimum eigenvalue of $n^{-1}X^TX$. We use the value (default) of $\gamma = 3$, so that only the covariates with nonzero coefficients are included in the calculation of c_*. Thus, the local convexity of the objective function will not be an issue for large λ, but may cease to hold as λ is reduced below critical value λ^*. Thus, the penalty is indexed by a regularization parameter λ, which controls the tradeoff between loss and penalty.

Since a grid of 100 values for λ that averages 10 iterations until convergence at each point, hence the algorithm calculates 1000 lasso paths to produce a single approximation to the MCP. The function `cv.ncvreg` is used for cross-validation and the `lambda.min` for obtaining the value of λ that gives minimum mean cross-validated error. To obtain the optimum number of variables and the best possible value of R–Squared, adjusted R–Squared and predicted R–Squared, the value lambda minimum are varied. Since MCP is a nonconvex penalty, on a large number of occasions they fail to converge. Hence, to ensure convergence, the smallest value for lambda, as a fraction of maximum lambda is 0.091. This is the smallest value lambda, for which all penalized coefficients become zero. The penalty applied to the model is "MCP" (the default). The tuning parameter gamma (γ) of this MCP penalty is three. For fair comparison, the same seed of 5226 are considered. The matrix of 144 observation and 986 covariates has 900 bad observations with value more than 100 and less than 0 (zero), which were replaced by the mean.

MCP allow the estimated coefficients to reach large values more quickly than the elastic net. In other words, MCP applies less shrinkage to the nonzero coefficients. The tuning parameter γ for the MCP estimates controls how fast the penalization rate goes to zero. The objective function is not locally convex in the shaded region, and hence the solutions are discontinuous and erratic. However, the solutions in the locally convex regions are continuous and stable.

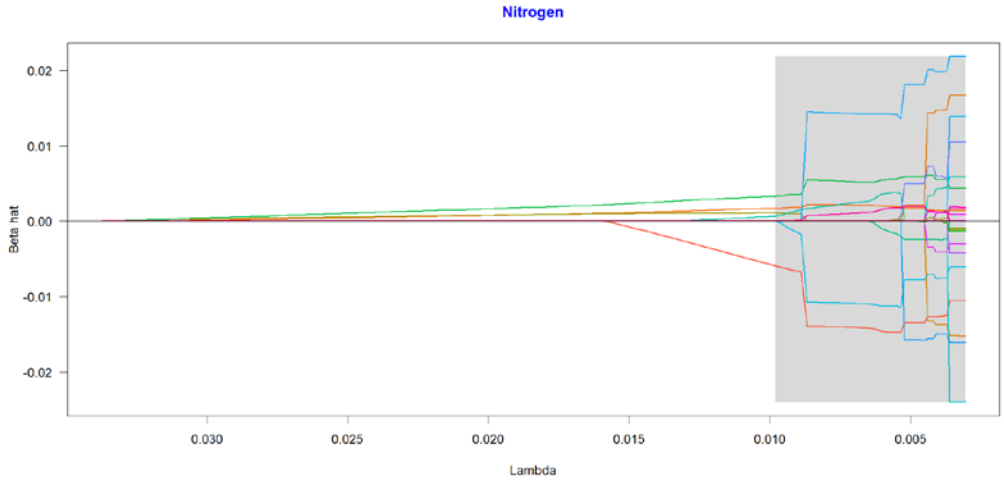


Figure 9.19: MCP Coefficient Paths for the response variable - Nitrogen

The Figure 9.19, Figure 9.22, Figure 9.25, Figure 9.28, Figure 9.31 and Figure 9.34 display the MCP coefficient paths for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. Notice how fast the penalization rate goes to zero on selecting the tuning parameter gamma (γ) = 3 for the MCP estimates. The shaded region depicts areas that are not locally convex. The value of lambda (λ_{min}) for all nutrients except for Magnesium lies outside the shaded region; hence, their solutions are continuous and stable. Since the value of λ_{min} for Magnesium lies inside the shaded region, its solutions are discontinuous and erratic. Some variables and the value of lambda (λ_{min}), at which the variables enter the model have been tabulated in Table 9.2.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Number of variables entering model	8	7	9	6	13	5
Lambda minimum (λ_{min})	0.01	605	680	120	2.7	0.6
λ_{min} lies outside the shaded region	Yes	Yes	Yes	No	Yes	Yes

Table 9.2: MCP coefficient paths of response variable of the grapevine dataset

Typically, one would carry out cross-validation to assess the predictive accuracy of the model at various values of λ.

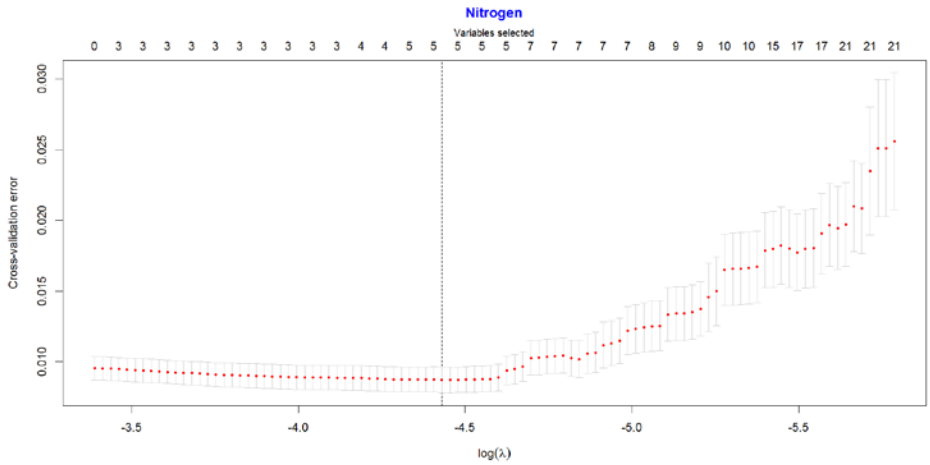
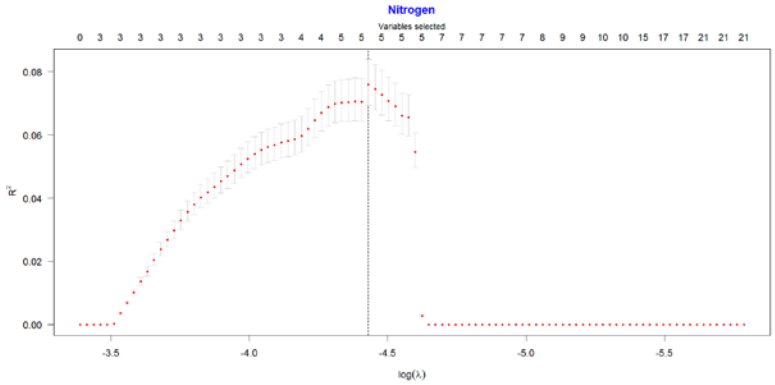


Figure 9.20: MSE and log (λ) using MCP for the response variable - Nitrogen

The Figure 9.20, Figure 9.23, Figure 9.26, Figure9.29, Figure 9.32 and Figure 9.35 shows that the cross-validation error for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The value of the lambda minimum (λ_{min}) and a log of lambda minimum corresponding to the minimum value of cross-validation error have been tabulated in Table 9.3. Some variables corresponding to the minimum value of cross-validation error and number of statistically significant variables have been included in the response variable of grapevine dataset.

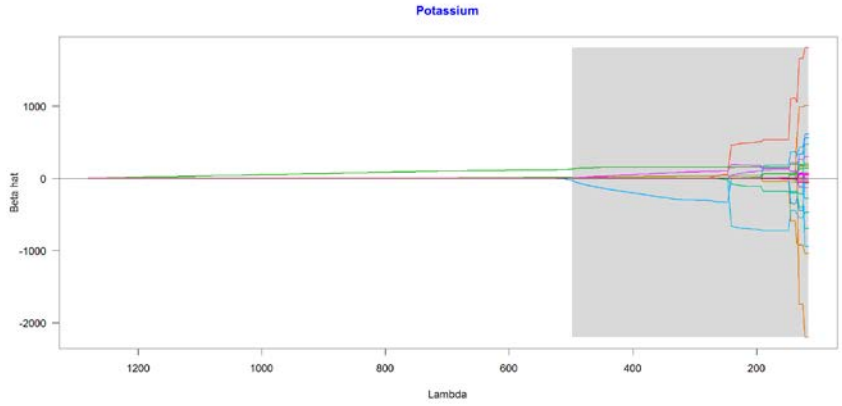
	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Lambda minimum (λ_{min})	0.01	604.87	680.37	119.74	2.67	0.63
Log Lambda minimum	-4.43	6.41	6.52	4.79	0.98	-0.46
No. of variables at min cross validation error	5	3	3	4	2	3
No. of significant variables	3	4	3	2	4	2

Table 9.3: Lambda values for response variable of the grapevine dataset using MCP

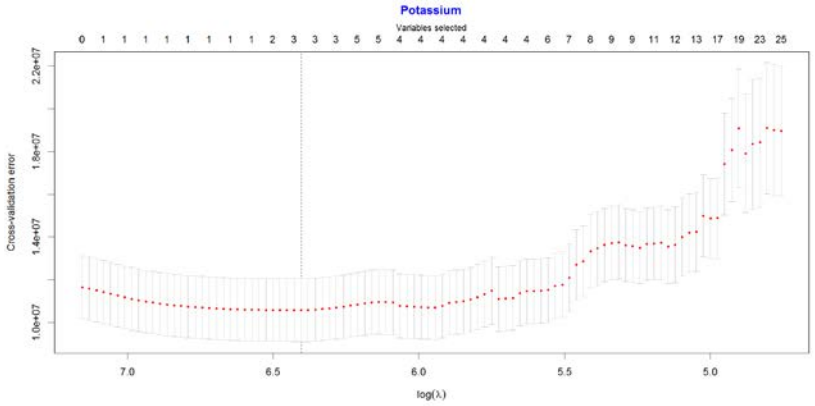


`Figure 9.21:` R-Squared and log (λ) using MCP for the response variable - Nitrogen

The `Figure 9.21` shows that the R-Squared value for nitrogen is maximum at the lambda (λ_{min}) of 0.01 and its log lambda value of -4.43. On either side, the value of R-Squared drops significantly. Even the maximum value of R-Squared indicates that the five variables explain about 8% of the variance.



`Figure 9.22:` MCP Coefficient Paths for the response variable - Potassium



`Figure 9.23:` MSE and log (λ) using MCP for the response variable - Potassium

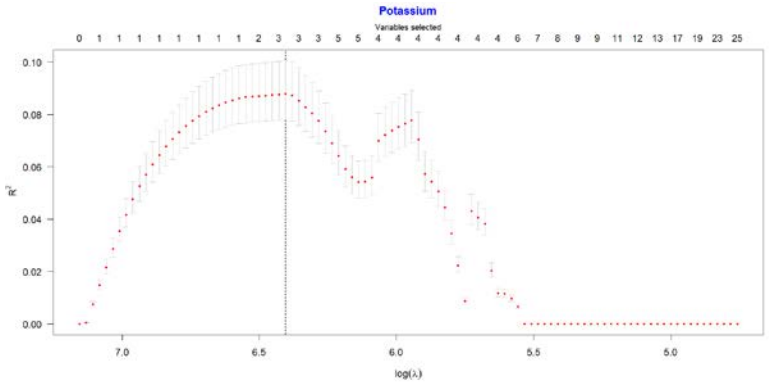


Figure 9.24: R-Squared and log (λ) using MCP for the response variable - Potassium

The Figure 9.24 shows that the R-Squared value for potassium is maximum at the lambda (λ_{min}) of 605 and its log lambda value of 6.4. On either side, the value of R-Squared drops significantly. Even the maximum value of R-Squared indicates that the three variables explain about 9% of the variance.

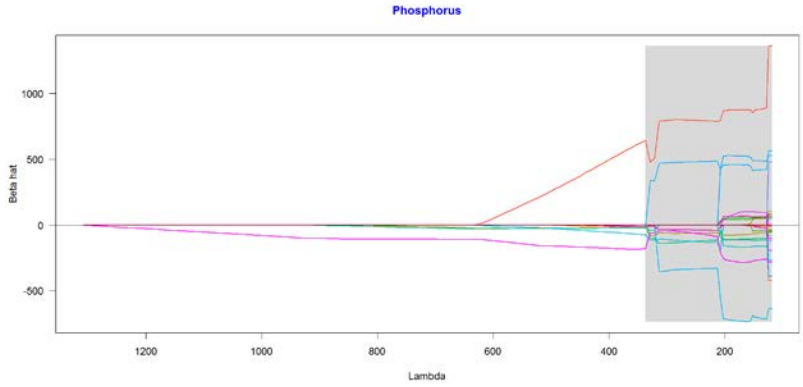


Figure 9.25: MCP Coefficient Paths for the response variable - Phosphorus

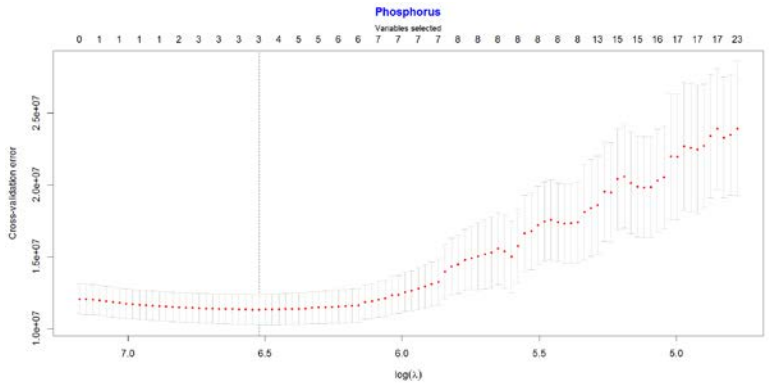


Figure 9.26: MSE and log (λ) using MCP for the response variable - Phosphorus

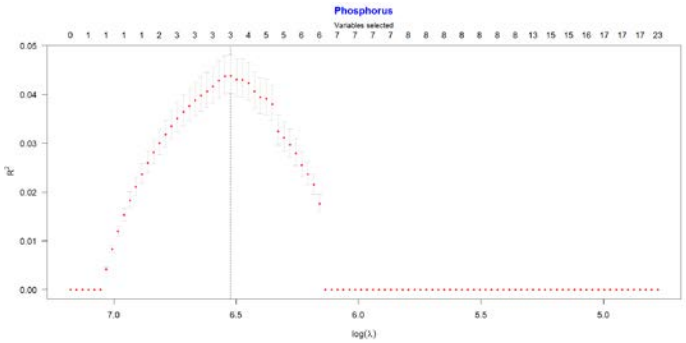


Figure 9.27: R-Squared and log (λ) using MCP for the response variable - Phosphorus

The Figure 9.27 shows that the R-Squared value for phosphorus is maximum at the lambda (λ_{min}) of 680 and its log lambda value of 6.52. On either side, the value of R-Squared drops significantly. Even the maximum value of R-Squared indicates that these three variables explain about 4.5% of the variance.

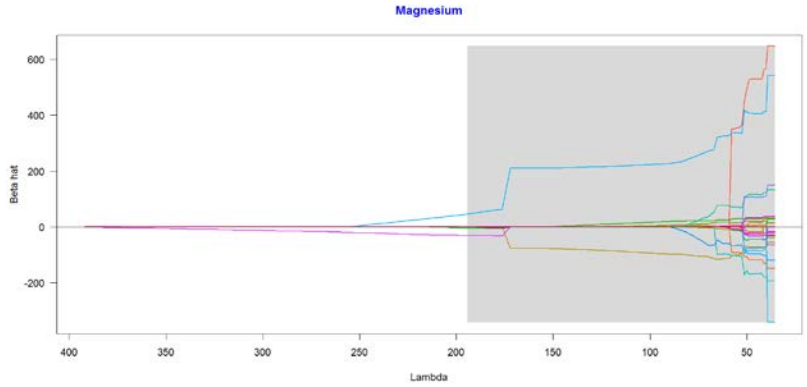


Figure 9.28: MCP Coefficient Paths for the response variable - Magnesium

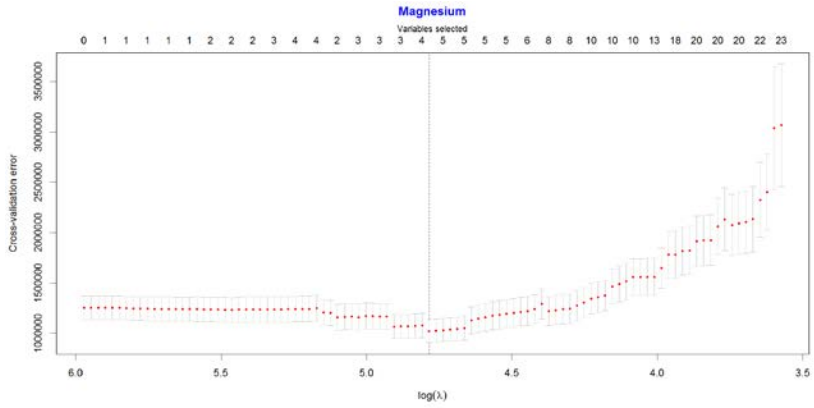


Figure 9.29: MSE and log (λ) using MCP for the response variable - Magnesium

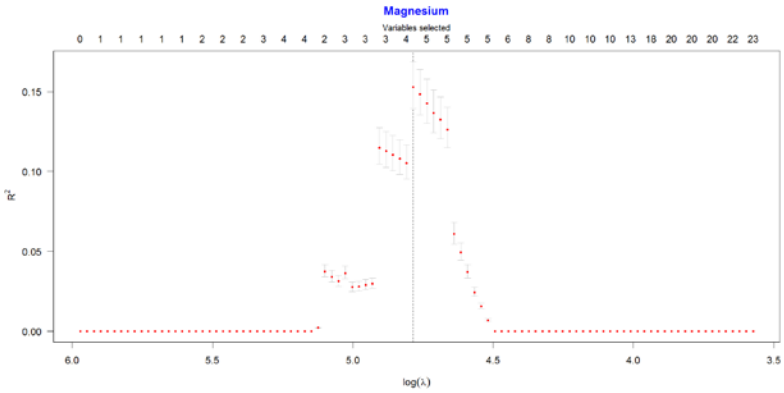


Figure 9.30: R-Squared and log (λ) using MCP for the response variable - Magnesium

The Figure 9.30 shows that the R-Squared value for magnesium is maximum at the lambda (λ_{min}) of 120 and its log lambda value of 4.8. On either side, the value of R-Squared drops significantly. Even the maximum value of R-Squared indicates that these four variables explain about 15% of the variance.

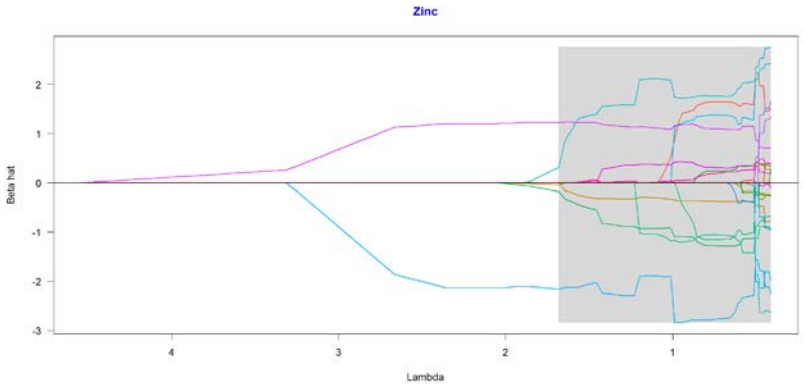


Figure 9.31: MCP Coefficient Paths for the response variable - Zinc

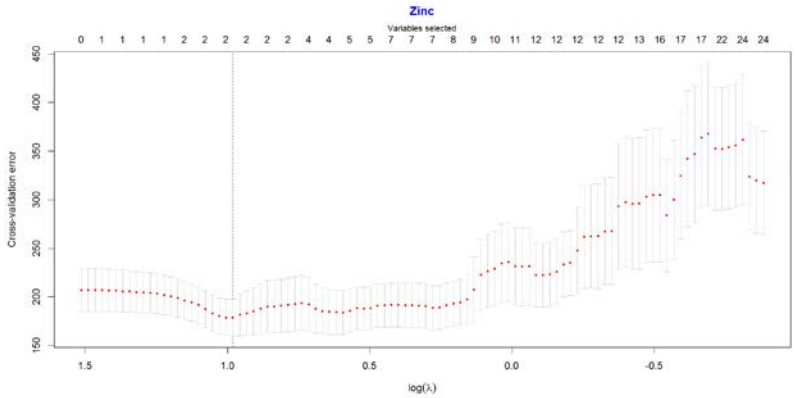


Figure 9.32: MSE and log (λ) using MCP for the response variable - Zinc

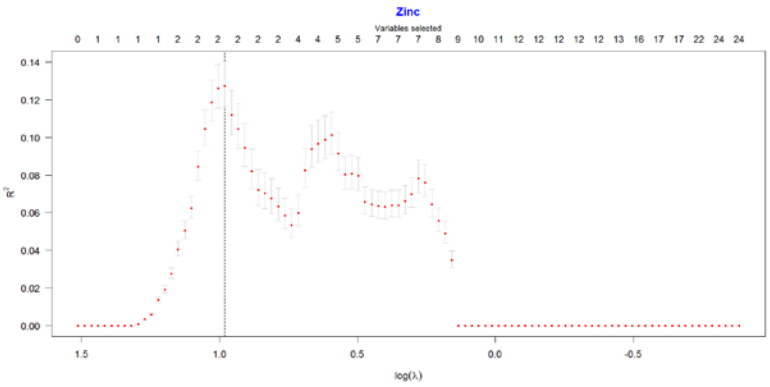


Figure 9.33: R-Squared and log (λ) using MCP for the response variable - Zinc

The Figure 9.33 shows that the R-Squared value for zinc is maximum at the lambda (λ_{min}) of 2.67 and its log lambda value of 0.98. On either side, the value of R-Squared drops significantly. Even the maximum value of R-Squared indicates that these three variables explain about 13% of the variance.

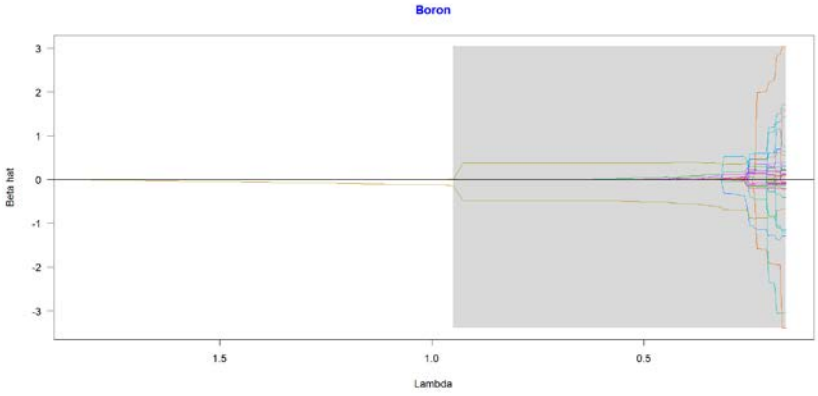


Figure 9.34: MCP Coefficient Paths for the response variable - Boron

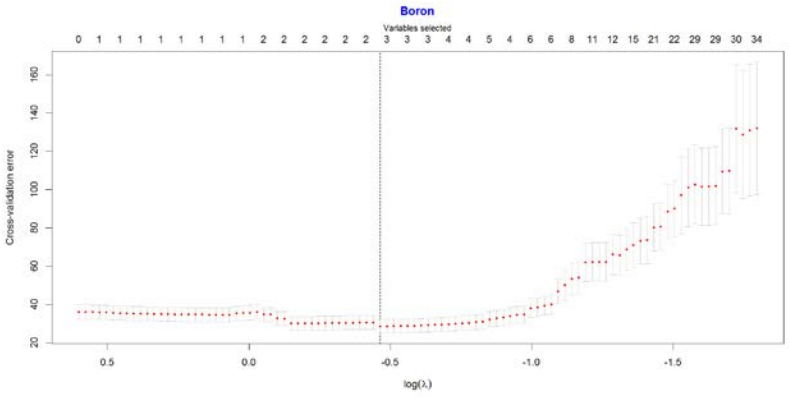


Figure 9.35: MSE and log (λ) using MCP for the response variable - Boron

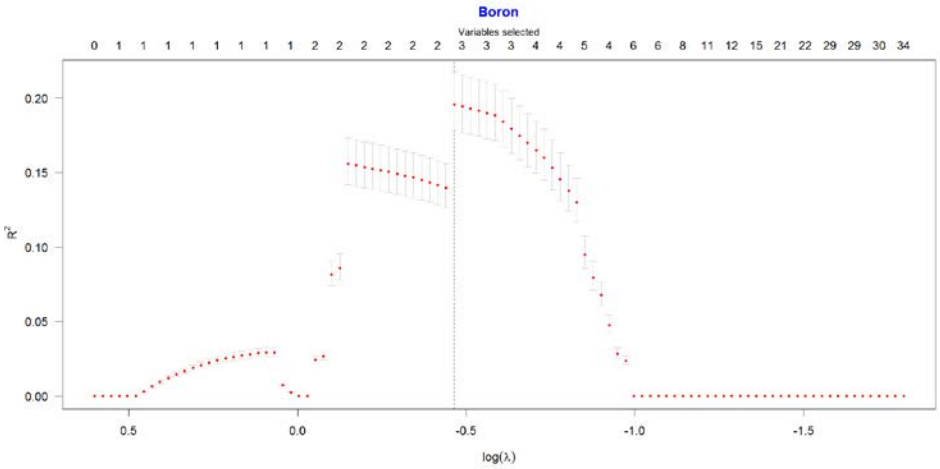


Figure 9.36: R-Squared and log (λ) using MCP for the response variable - Boron

The Figure 9.36 shows that the R-Squared value for boron is maximum at the lambda (λ_{min}) of 0.63 and its log lambda value of -0.46. On either side, the value of R-Squared drops significantly. Even the maximum value of R-Squared indicates that these three variables explain about 18% of the variance.

The value for R–Squared, adjusted R–Squared and predicted R–Squared for the six nutrients are given below:

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.squared	0.19	0.18	0.22	0.34	0.26	0.17
Adj.R.Squared	0.18	0.16	0.20	0.33	0.24	0.15
Pred.R.Squared	0.15	0.13	0.17	0.31	0.21	0.14

MCP performs well when there are many rather sparse groups of predictors, i.e. when the underlying model exhibits less grouping of predictors. MCP suffers when the non-zero coefficients are clustered into tight groups. MCP makes insufficient use of the grouping information and hence, selects too few cluster. Since the grapevine dataset is clustered into tight groups, the sparse solution due to MCP selects a smaller number of predictors than desired. The process mentioned above has adversely affected the predictive ability of regression model based on Minimax Concave Penalty. The values of R–Squared are close to the adjusted R–Squared and predicted R–Squared values. However, they are lower than the linear regression using elastic net penalty.

9.5 Iterative Sure Independence Screening using the SIS (R package)

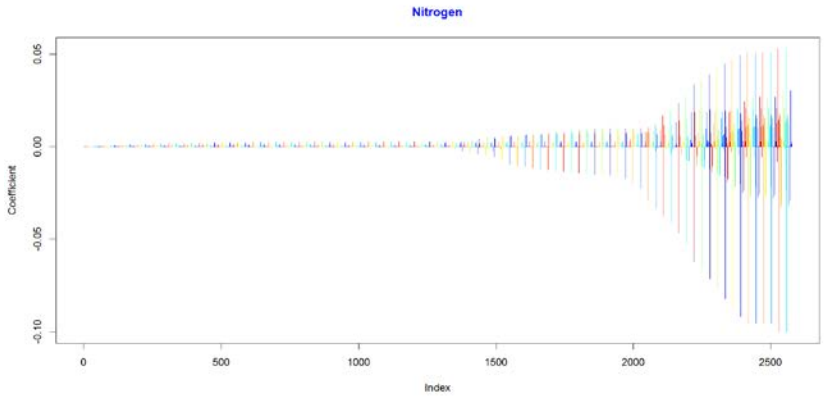
Third, to carry out the Sure Independence Screening (SIS) variable selection procedure, we initially fit marginal versions of models with component-wise covariates. To avoid the numerical instability associated with high-dimensional estimation problems, we need to compute component wise estimators and implement modularly. The SIS package then ranks the importance of features according to the magnitude of their marginal regression coefficients, excluding the intercept in the case of GLM. Therefore, a set of variables is given below.

$$\widehat{\mathcal{M}}_{\delta_n} = \{1 \leq j \leq p: |\hat{\beta}_j^M| \geq \delta_n\}$$

Where δ_n is a threshold value chosen so that top-ranked covariates are picked, so that dimensionality is reduced from ultrahigh to below the sample size, we consider d = [n/log n]. Improvement of finite sample performance using SIS, variable selection, and parameter estimation can be simultaneously achieved via penalized likelihood estimation, using the joint information of the covariates in $\widehat{\mathcal{M}}_{\delta_n}$ (Saldana & Feng, 2016).

Iterative Sure Independence Screening (ISIS) fits the regression model using the R packages `ncvreg` and `glmnet` for regularized log likelihood for the variables selection by ISIS. In this case, “lasso” is selected as the penalty for the regularized likelihood for the sub-problems and “AIC” for tuning the regularization parameter of the penalized likelihood for the sub-problems and the final model selected by ISIS. By nature of their marginal approach, sure independence screening procedures have massive false selection rates, $\mathcal{M}_*^c$ are selected after the screening steps. In order to reduce the false selection rate, (Saldana & Feng, 2016) suggested the idea of sample splitting. Without loss of generality, the SIS package has randomly split the sample into two halves, used random permutation, and cross- validation sampling of training and test sets (Fan et al., 2016). Taking advantage of the fast cyclical coordinate descent algorithms developed in the packages `glmnet` (J Friedman et al., 2013) and `ncvreg` (Breheny & Breheny, 2016), for convex and nonconvex penalty functions, respectively, we were able to efficiently perform the moderate scale penalized pseudo-likelihood steps from the ISIS procedure. This variable selection technique outperforms direct use of `glmnet` and `ncvreg` in terms of both computational time and estimation error.

The function `SIS` initially makes 20 attempts to split the complete sample. After that, it tries a more conservative variable screening approach with a data-driven threshold for marginal screening.



`Figure 9.37:` Plot of beta coefficients for the response variable - Nitrogen

The `Figure 9.37` to `Figure 9.42` displays the sparse matrix of the beta coefficients for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. After a certain number of iterations for screening, the sure independence screening method selects significant predictive variables for the response variables of the grapevine dataset. The coefficient of the remaining predictive variables is reduced to zero. Some iterations and significant variables for response variables of the grapevine have been tabulated in `Table 9.4`.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
No. of iterations	3	4	3	2	2	3
No. of significant variables	1	2	11	4	8	9

`Table 9.4:` Iterations and significant variables for response variables of the grapevine dataset using SIS

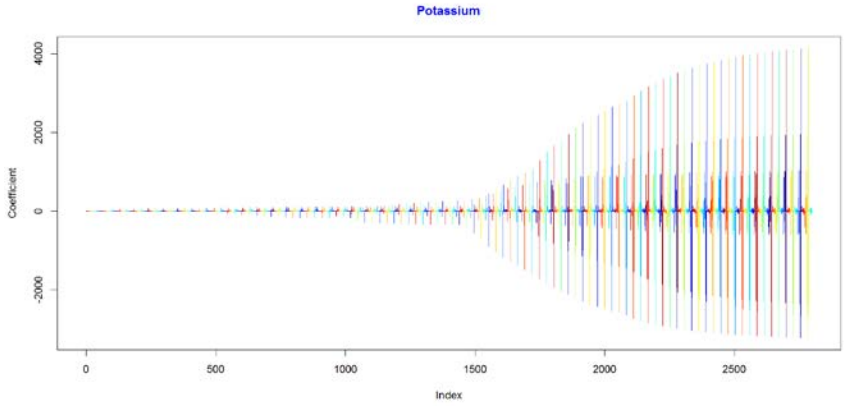


Figure 9.38: Plot of beta coefficients for the response variable - Potassium

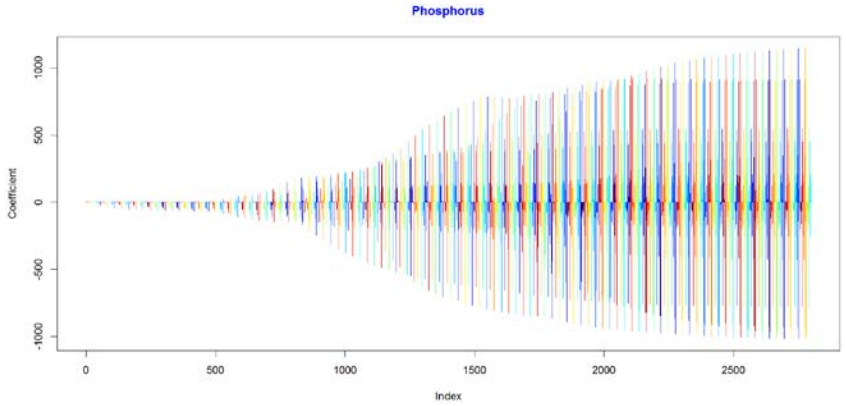


Figure 9.39: Plot of beta coefficients for the response variable - Phosphorus

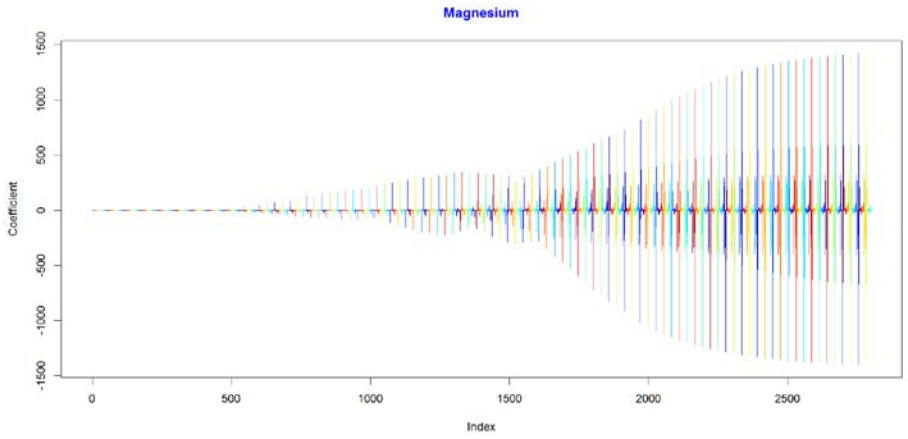


Figure 9.40: Plot of beta coefficients for the response variable - Magnesium

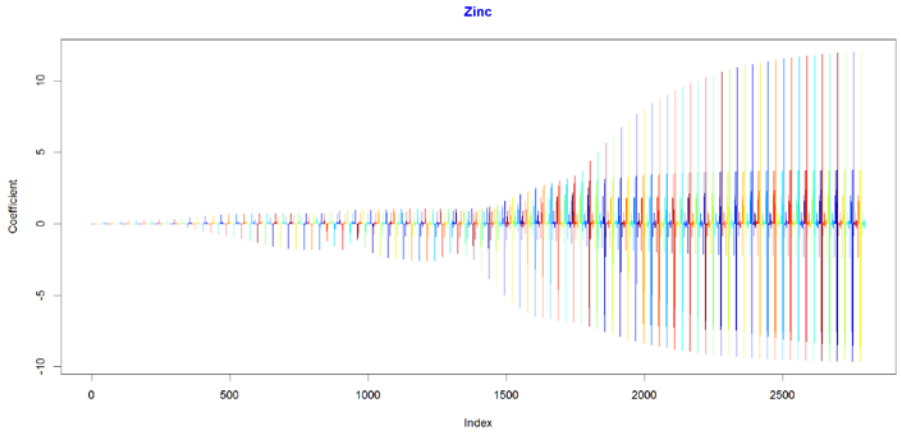


Figure 9.41: Plot of beta coefficients for the response variable - Zinc

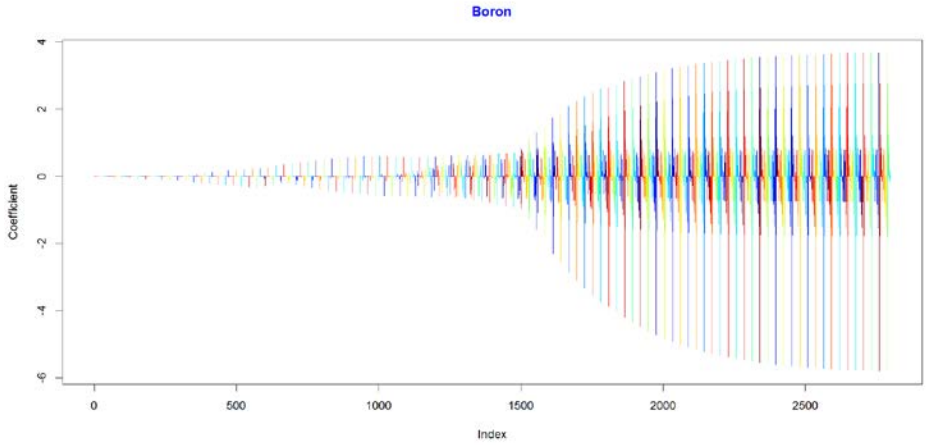


Figure 9.42: Plot of beta coefficients for the response variable - Boron

The values of R–Squared, adjusted R–Squared and predicted R–Squared for the six nutrients using iterative sure independence screening are given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.squared	0.04	0.22	0.41	0.33	0.41	0.38
Adj.R.Squared	0.04	0.2	0.36	0.31	0.38	0.34
Pred.R.Squared	0.02	0.19	0.29	0.28	0.31	0.31

SIS computes, component wise estimators using the method of AIC for tuning the regularization parameter of the penalized likelihood Lasso. This procedure iteratively performs variable selection to recruit a small number of predictors and computes residuals based on the model fitted using these recruited predictors. Then these residuals are used as the working response variable to continue recruiting new predictors.

All the variables are selected within 2 - 3 iterations except potassium, where four iterations were required. Except for nitrogen, the values of R–Squared, adjusted R–Squared and predicted R–Squared for the other nutrients are either comparable or better than the value obtained using the nonconvex penalty of MCP in the package `ncvreg`. However, these values are lower than the one obtained using convex penalty of the elastic net in the package `glmnet`.

9.6 Functional Data Analysis using package `fda.usc`

Since, the value of reflectance has been taken for wavelengths, spread between 334 nanometers (nm) and 2510 nanometers separated by 1.5 to 2.7 nm. Hence, we can regard the spectral reflectance data measured along the continuum of wavelength as a single entity.

Then using the function `fdata` from the `fda.usc` package, we can convert the data (predictors) object of class "matrix" or "data.frame" to an object of class "fdata" by basis of smoothing, where [1,986] is the range of discretization points. This representation, which implicitly assumes a ℓ_2 space, is not related to the information of the response variable. In other words, the vertical shift of these curves has no special relation with the nutrients. Since predictors are a non-periodic functional data, we can use spline functions for approximation, which combines the fast computation of polynomials with substantially greater flexibility and a modest number of basis functions. Then `fregre.basis` function in the `fda.usc` package computes functional regression between functional explanatory variable $X(t)$ and scalar response Y is one of the six nutrients using B-spline (default) basis representation.

$$Y = \langle X, \beta \rangle + \epsilon = \int_T X(t)\beta(t)dt + \epsilon$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of the covariates of the reflectance value of the grapevine data on ℓ_2 space, and ϵ are random errors with mean zero, finite variance σ^2 and $E[X(t)\epsilon] = 0$ (Bande et al., 2016).

This function allows covariates of class "fdata," "matrix," or "data.frame" and gives default values to arguments basis.x and basis.b for representation by functional data $X(t)$ and the functional

parameter $\beta(t)$, respectively. We do not consider any roughness penalty (λ) for this functional data. In addition, the function fregre.basis.cv uses the validation criterion to estimate the number of basis elements and/or the penalized parameter (λ) that best predicts the response.

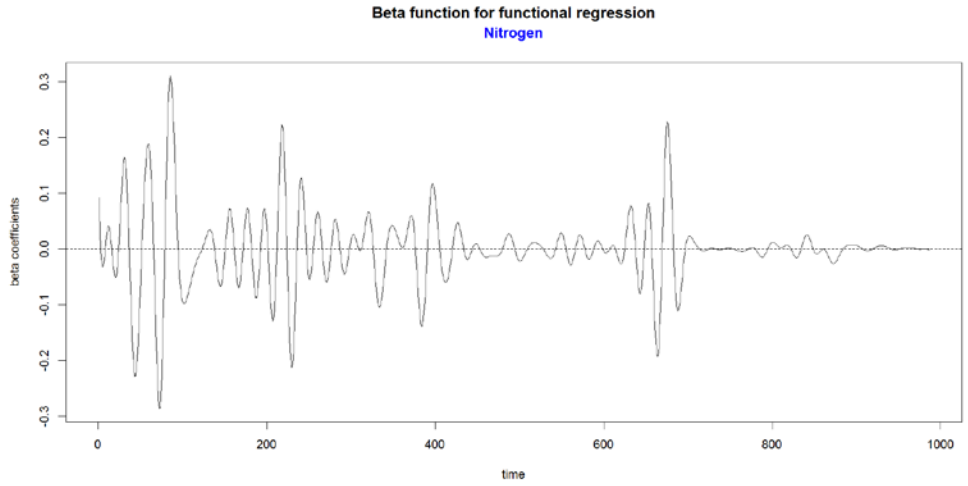


Figure 9.43: Beta coefficient of response variable - Nitrogen for Functional Regression

The Figure 9.43, Figure 9.47, Figure 9.51, Figure 9.55, Figure 9.59 and Figure 9.63 displays the plot of the beta coefficients for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. 98-basis functions (K) with zero roughness penalty were used to smooth the data. B-spline basis representation was used to compute the functional regression between functional explanatory variable (spectral reflectance) X (t) and scalar response variables of grapevine dataset. Since the number of basis functions (K) is not substantially smaller than the number of observations (n) of 144, the regression approach tends to overfit the data.

In spline smoothing, as in other smoothing methods, the mean squared error (MSE), is one way of capturing the quality of the estimate. For imposing smoothness on the estimated curve, MSE is reduced by sacrificing some bias to reduce sampling variance. Since the estimates are expected to vary gently from one value to another, we are effectively borrowing information from neighboring data values, thereby expressing our faith in the regularity of the underlying function, x, that we are trying to estimate. This pooling of information is what makes our estimated curve more stable, at the cost of some increase in bias (J. Ramsay & Silverman, 2005).

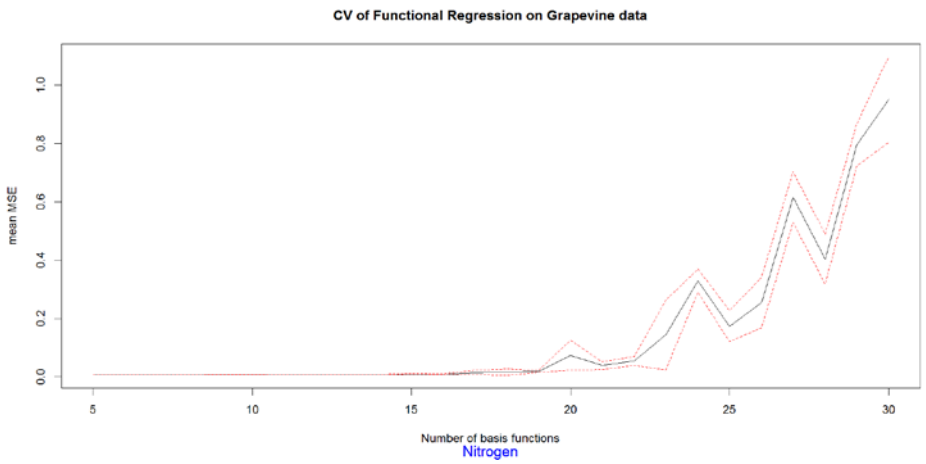


Figure 9.44: CV of Functional Regression for response variable - Nitrogen

The Figure 9.44 shows cross-validation of functional regression of grapevine data for the response variable nitrogen, based on ten iterations. Based on minimum mean MSE the number of basis function appears to be between 5 and 16.

It is desirable to have a lower-dimensional B-spline basis defined by some appropriate more limited knot sequence, τ, provided there remain sufficient flexibility to capture the features of interest.

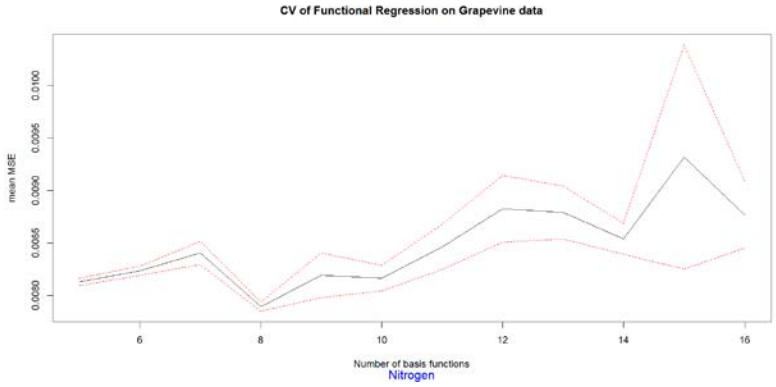


Figure 9.45: CV of Functional Regression for response variable - Nitrogen

The Figure 9.45 shows cross-validation of functional regression of grapevine data for the number of basis functions between 5 and 16, based on 30 iterations. Based on minimum mean MSE the number of basis function appears to be 8.

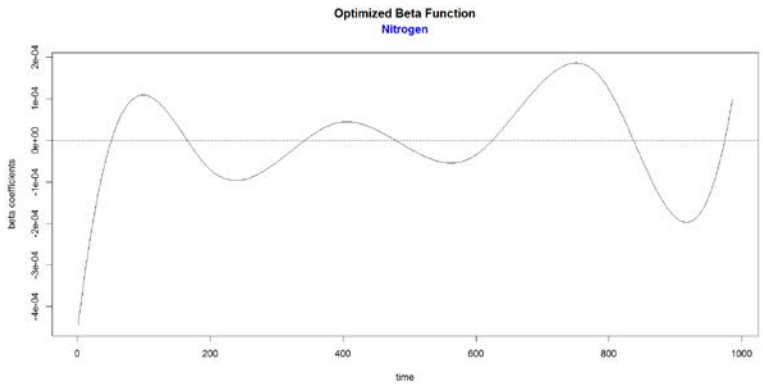


Figure 9.46: Optimized beta function for response variable - Nitrogen

The optimized beta function for nitrogen, based on eight basis functions, was obtained from Figure 9.45.

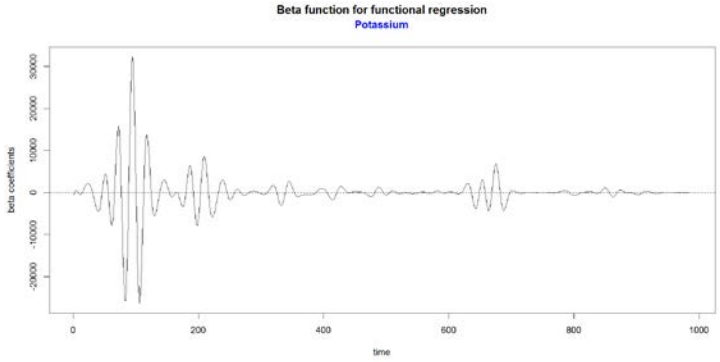


Figure 9.47: Beta coefficient of Functional Regression for response variable - Potassium

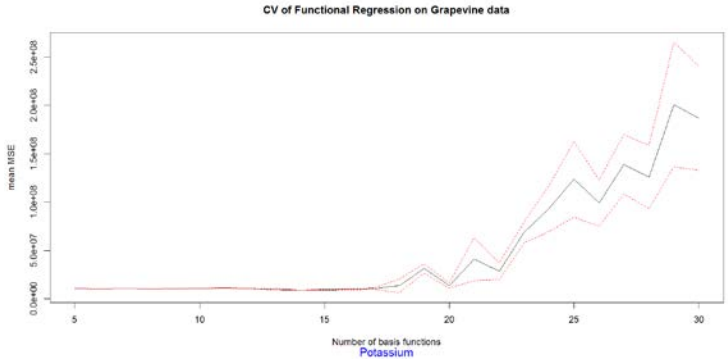


Figure 9.48: CV of Functional Regression for response variable - Potassium

The Figure 9.48 shows cross-validation of functional regression of grapevine data for the response variable potassium, based on ten iterations. Based on minimum mean MSE, the number of basis functions appears to be between 5 and 15.

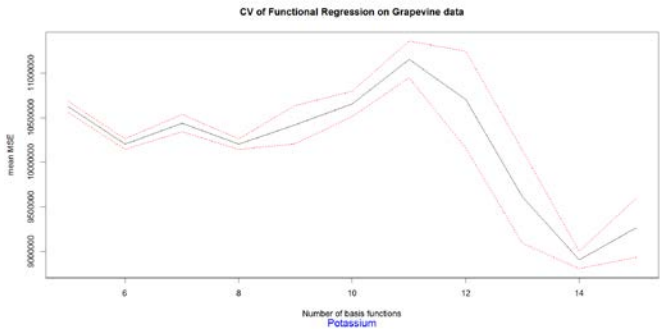


Figure 9.49: CV of Functional Regression for response variable - Potassium

The Figure 9.49 shows cross-validation of functional regression of grapevine data for some basis functions between 5 and 15 based on 30 iterations. Based on minimum mean MSE, the number of basis functions appears to be 14.

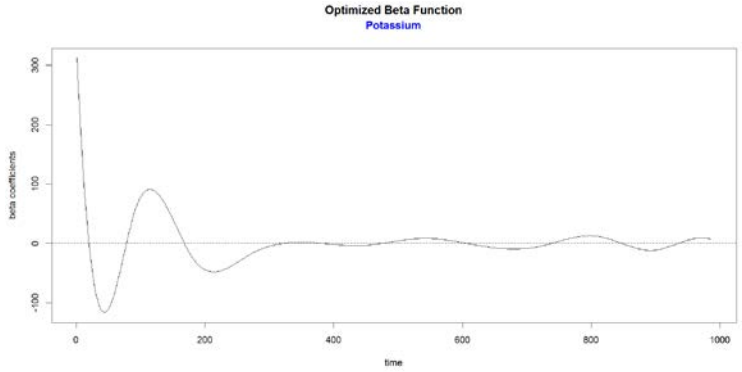


Figure 9.50: Optimized beta function for response variable - Potassium

The optimized beta function for Potassium based on 14, basis function obtained from the Figure 9.49.

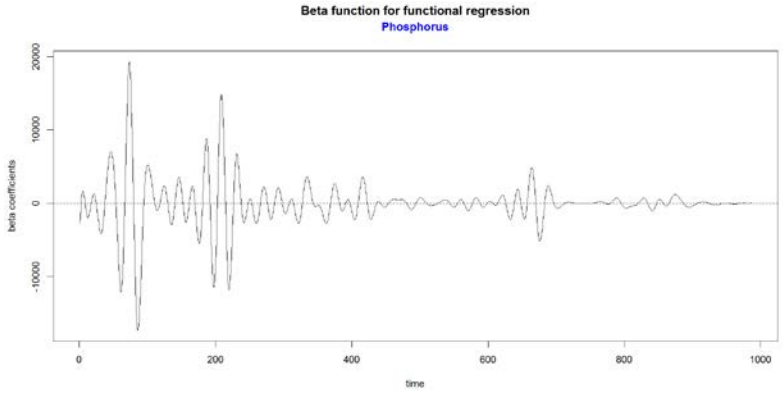


Figure 9.51: Beta coefficient of Functional Regression for response variable - Phosphorus

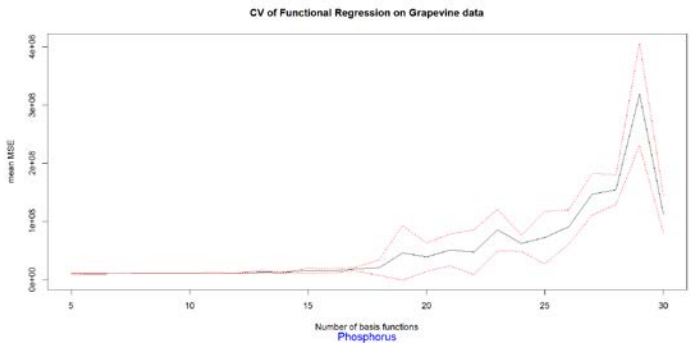


Figure 9.52: CV of Functional Regression for response variable - Phosphorus

The Figure 9.52 shows cross-validation of functional regression of grapevine data for the response variable phosphorus, based on ten iterations. Based on minimum mean MSE, the number of basis functions appears to be between 5 and 15.

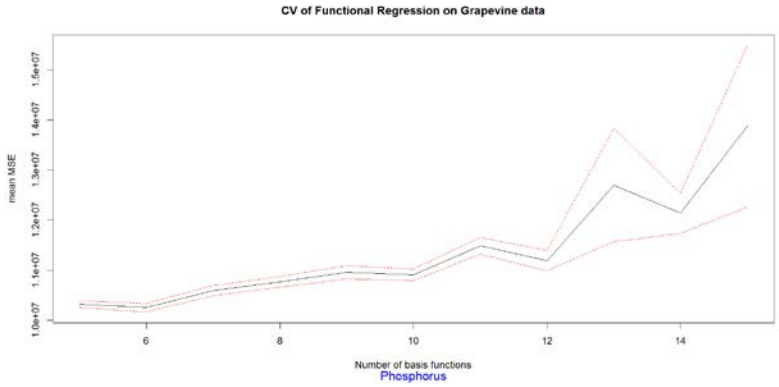


Figure 9.53: CV of Functional Regression for response variable - Phosphorus

The Figure 9.53 shows cross-validation of functional regression of grapevine data for some basis functions between 5 and 15, based on 30 iterations. Based on minimum mean MSE, the number of basis functions appears to be six.

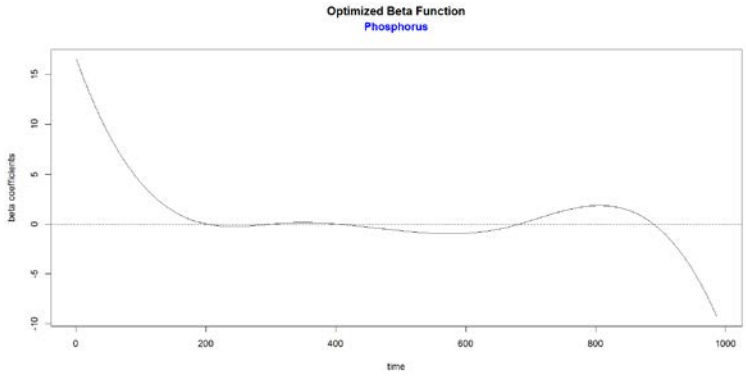


Figure 9.54: Optimized beta function for response variable - Phosphorus

The optimized beta function for Phosphorus-based on six, basis function obtained from the Figure 5.53.

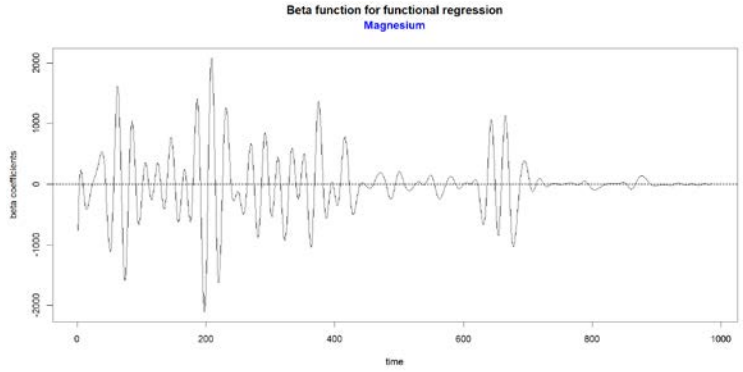


Figure 9.55: Beta coefficient for Functional Regression of response variable - Magnesium

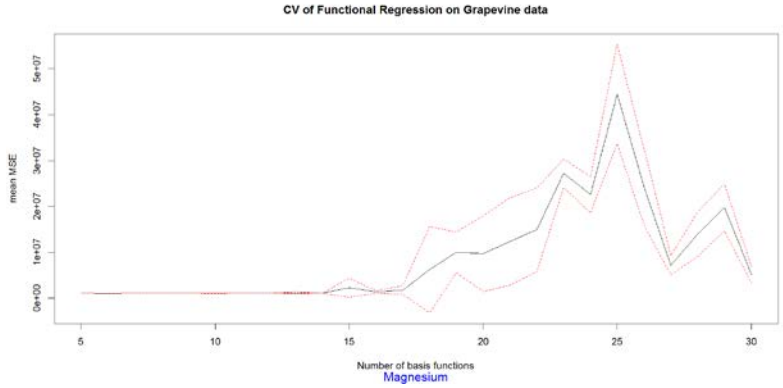


Figure 9.56: CV of Functional Regression for response variable - Magnesium

The Figure 9.56 shows cross-validation of functional regression of grapevine data for the response variable magnesium, based on ten iterations. Based on minimum mean MSE, the number of basis functions appears to be between 5 and 14.

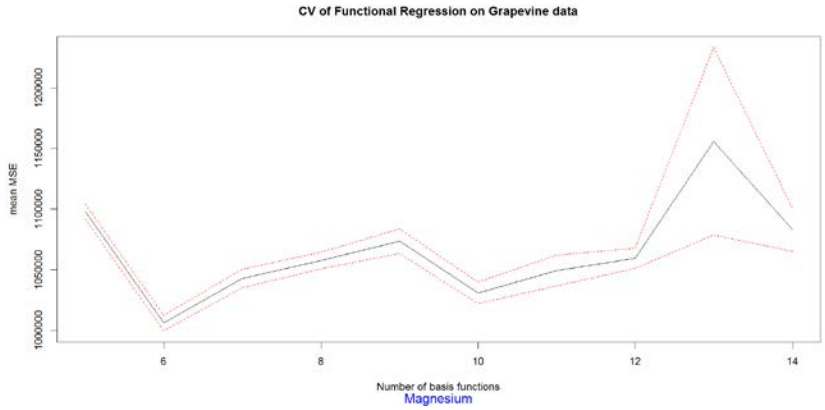


Figure 9.57: CV of Functional Regression for response variable - Magnesium

The Figure 9.57 shows cross-validation of functional regression of grapevine data for some basis functions between 5 and 14, based on 30 iterations. Based on minimum mean MSE, the number of basis function appears to be six.

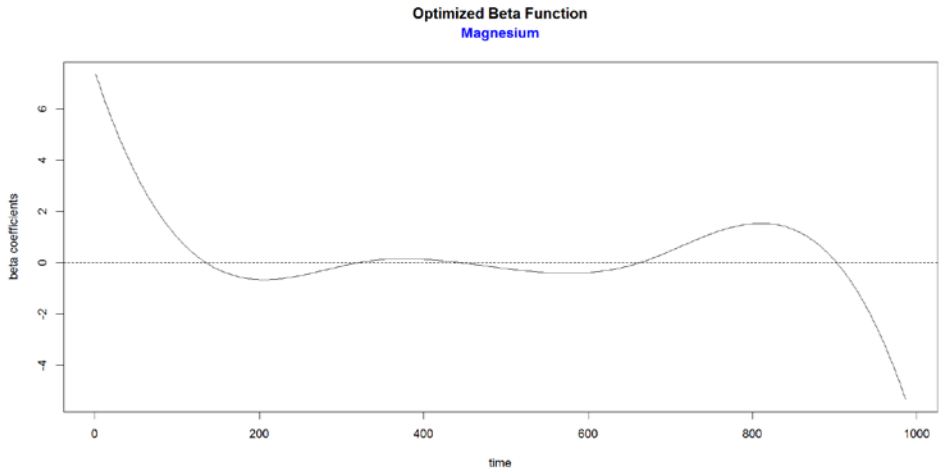


Figure 9.58: Optimized beta function for response variable - Magnesium

The optimized beta function for magnesium based on six, basis function obtained from the Figure 9.57.

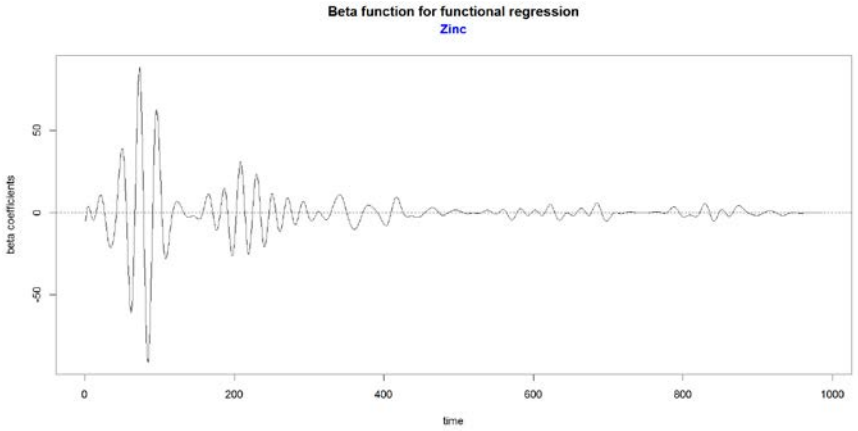
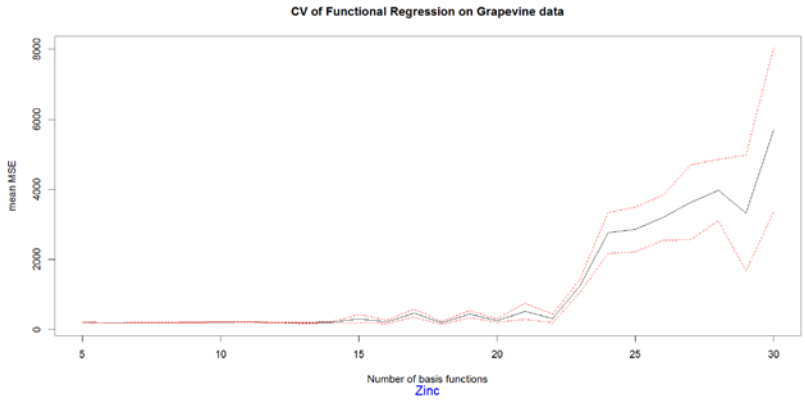
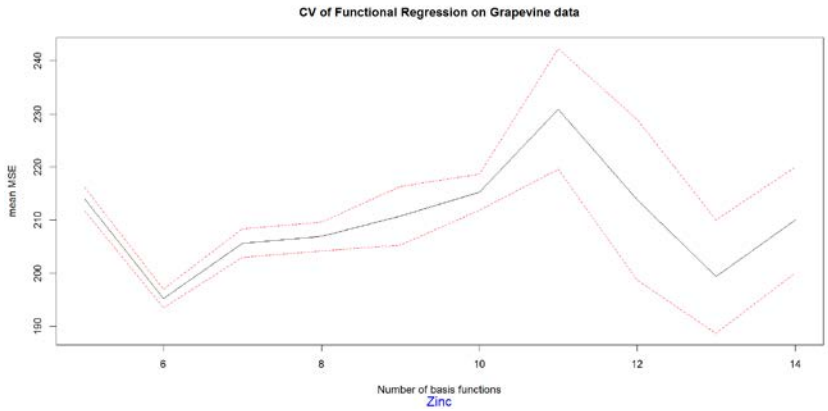


Figure 9.59: Beta coefficient of Functional Regression for response variable - Zinc



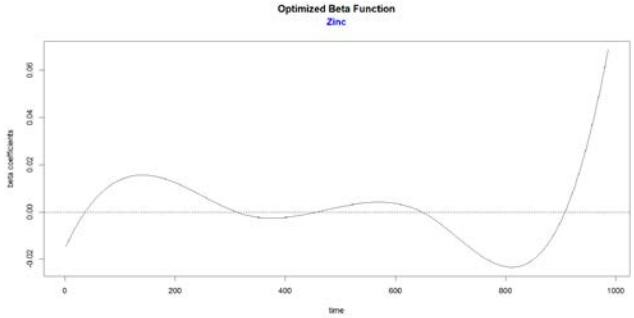
`Figure 9.60:` CV of Functional Regression for response variable - Zinc

The `Figure 9.60` shows cross-validation of functional regression of grapevine data for the response variable zinc, based on ten iterations. Based on minimum mean MSE, the number of basis functions appears to be between 5 and 14.



`Figure 9.61:` CV of Functional Regression for response variable - Zinc

The `Figure 9.61` shows cross-validation of functional regression of grapevine data for some basis functions between 5 and 14, based on 30 iterations. Based on minimum mean MSE, the number of basis functions appears to be six.



`Figure 9.62:` Optimized beta function for response variable - Zinc

The optimized beta function for zinc based on six, basis functions obtained from the Figure 9.61.

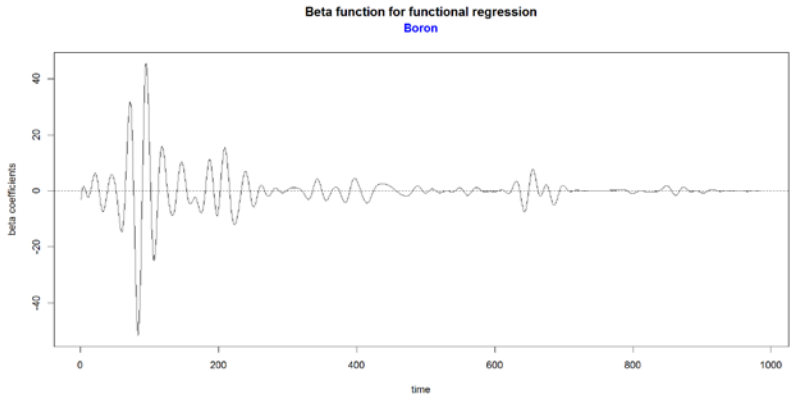


Figure 9.63: Beta coefficient of Functional Regression for response variable - Boron

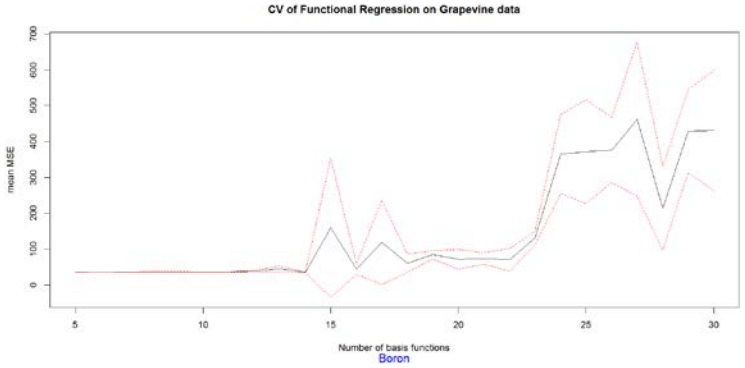


Figure 9.64: CV of Functional Regression for response variable - Boron

The Figure 9.64 shows cross-validation of functional regression of grapevine data for the response variable boron, based on ten iterations. Based on minimum mean MSE, the number of basis functions appears to be between 5 and 14.

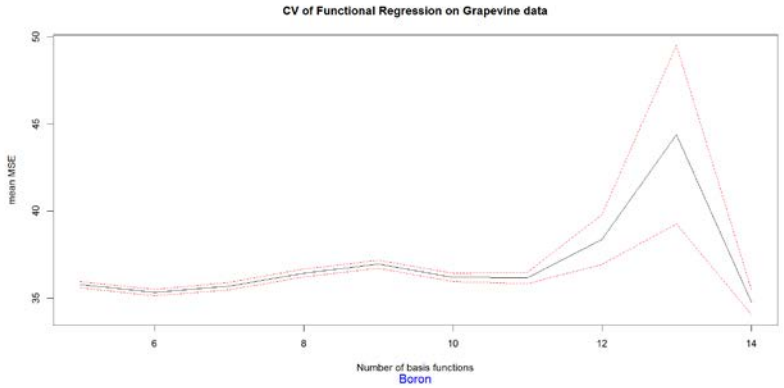


Figure 9.65: CV of Functional Regression for response variable - Boron

The Figure 9.65 shows cross-validation of functional regression of grapevine data for some basis functions between 5 and 14, based on 30 iterations. Based on minimum mean MSE, the number of basis functions appears to be 14.

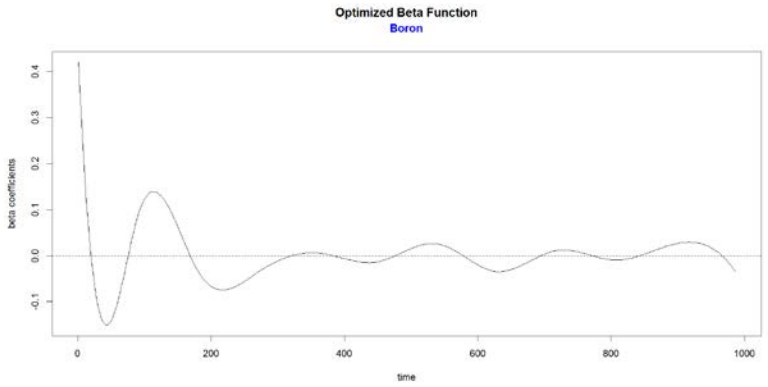


Figure 9.66: Optimized beta function for response variable - Boron

The optimized beta function for Boron based on 14, basis function obtained from the Figure 9.65.

By default 98 (10%) basis function is selected by the function fregre.basis , which can be verified using function summary.fregre.fd(). It is worth mentioning that only 10, 6, 6, 5, 11 and 5 basis functions are statistically significant for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. By carrying out cross validation, we can determine that the optimal number of basis function for nitrogen, potassium, phosphorus, magnesium, zinc and boron as 31, 13, 7, 29, 23, and 21, respectively. The fregre.basis.cv () uses validation criterion, which is defined to estimate the number of basis elements and/or the penalized parameter (λ) that best predicts the response. However, even these basis functions appears to be high, hence the optimal number of basis function was obtained by plotting mean MSE on y-axis against the number of basis functions on x-axis. The number of basis functions that gives the minimum mean MSE were chosen for further study. The values of R-Squared, adjusted R-Squared and predicted R-Squared based on the number of basis function obtained against the minimum value of mean MSE are given below.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.squared	0.25	0.37	0.23	0.25	0.19	0.30
Adj.R.Squared	0.20	0.31	0.19	0.22	0.16	0.22
Pred.R.Squared	0.17	0.25	0.14	0.17	0.06	0.07

We notice that the best values of R-squared, adjusted R-squared, and predicted R-squared could be achieved by using a generalized linear model via penalized maximum likelihood in the package `glmnet`. Selection of lambda was made using 10-fold cross-validation, based on mean squared error criterion. Hence, we will use the generalized linear model via penalized maximum likelihood in the package `glmnet`, for the rest of our calculation and discussion.

Chapter 10

Problem associated with Multivariate Dataset

10.1 Introduction

In this chapter, we study the grapevine data for leaf-level reflectance and petiole-level chemical analysis of the Riesling variety, taken during the bloom period of growth from the view angle directly over the individual grape leaves. In the last chapter, we have noticed that the best results for the value of R-squared, adjusted R-squared, and predicted R-squared values were obtained using the elastic net regularization path for fitting the generalized linear regression paths, by maximizing the appropriately penalized log-likelihood in the package `glmnet`.

Lambda min ratio is the smallest value for lambda, as a fraction of the maximum value of lambda. It is also the lowest value for which all coefficients are zero. The lambda min is the value of lambda that gives minimum mean cross-validated error - a vector of length (lambda). In this chapter, the comparative study of different values of lambda min ratio and lambda min has was examined, based on generalized linear model via penalized maximum likelihood. For a fair comparison, same parameters of seed= 5223, and `alpha`= 0.92 were selected. Since the predictors are highly correlated, slight variation in the value of Lambda min ratio and Lambda min, the selection of predictor variables change, impacting the values of R-squared, adjusted R-squared and predicted R-squared.

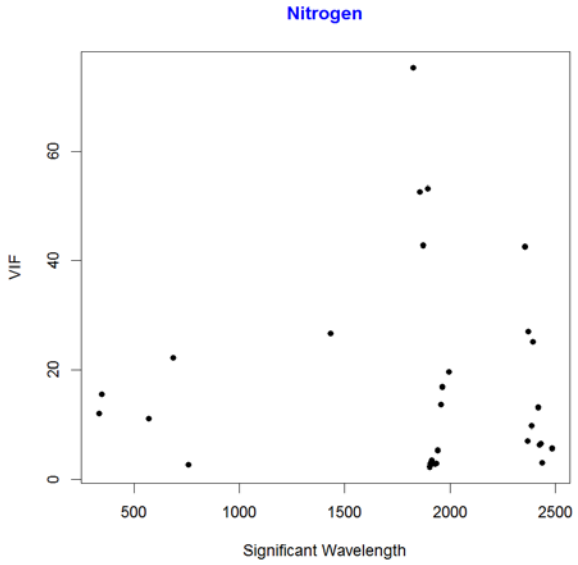
10.2 Value of lambda.min and lambda.min.ratio as 0.004

With the selection of lambda.min and lambda.min.ratio as 0.004, to calculate the optimum value, we got the following values of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.80	0.57	0.63	0.66	0.74	0.56
Adj.R.Squared	0.74	0.50	0.57	0.60	0.67	0.49
Pred.R.Squared	0.68	0.40	0.48	0.47	0.59	0.38

Significant Wavelength (Nitrogen; nm): 334.3, 347.8, 569.9, 684.6, 756.9, 1434.4, 1826.4, 1858.4, 1872.6, 1893.8, 1903.8, 1906.6, 1912.2, 1928.9, 1934.5, 1942.8, 1956.6, 1962.1, 1994.8, 2355.2, 2368.8, 2371, 2386.7, 2393.3, 2419.7, 2426.2, 2430.5, 2439.1, 2483.6

Variance Inflation Factors (VIF): 12.1, 15.58, 11.11, 22.26, 2.63, 26.67, 75.41, 52.66, 42.87, 53.26, 2.21, 2.84, 3.48, 2.78, 2.81, 5.34, 13.67, 16.97, 19.64, 42.63, 6.99, 27.03, 9.72, 25.12, 13.11, 6.31, 6.5, 3.04, 5.63



`Figure 10.1:` Scatterplot of VIF against Wavelength - Nitrogen

`Figure 10.1` to `Figure 10.6` displays the scatter plot of VIF against the wavelengths for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The concentration of significant wavelengths for the response variables for the grapevine dataset can be seen to have a VIF around 10. Certain significant wavelengths have high VIF; however, the median VIF of significant predictors (wavelength) has been tabulated in `Table 10.1`.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Median of VIF	Around 10	Around 12	Around 12	Around 7	Around 8	Around 7

`Table 10.1:` Median VIF of significant predictors for lambda.min of 0.004

Significant Wavelength (Potassium; nm):334.3, 338.8, 341.8, 515.1, 646.9, 867.9, 1348.1,1419.4, 1862, 1869.1, 1915, 1928.9, 1951.1, 1962.1, 1989.3, 2323.2, 2382.2, 2393.3, 2487.8, 2506.4

Variance Inflation Factors (VIF): 9.96, 20.58, 26.02, 68.94, 51.73, 1.85, 37.03, 9.87, 30.85, 30.67, 2.58, 3.32, 3.37, 4.29, 11.53, 38.08, 15.97, 16.54, 2.71, 2.36

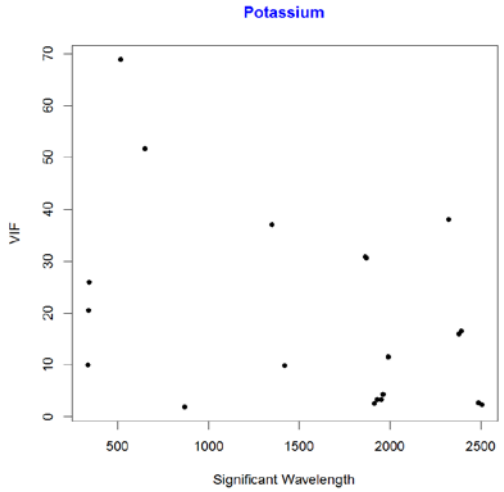


Figure 10.2: Scatterplot of VIF against Wavelength - Potassium

Significant Wavelength (Phosphorus; nm): 334.3, 359.7, 925.6, 1438.2, 1903.8, 2323.2, 2353, 2386.7, 2410.9, 2426.2, 2437, 2439.1, 2441.3, 2458.4, 2473.2, 2500.3, 1909.4, 1928.9

Variance Inflation Factors (VIF): 9.02, 8.76, 1.52, 15.11, 1.9, 22.89, 22.92, 7.17, 2.61, 5.65, 7.49, 2.64, 9.1, 5.69, 3.96, 2.79, 1.97, 2.27

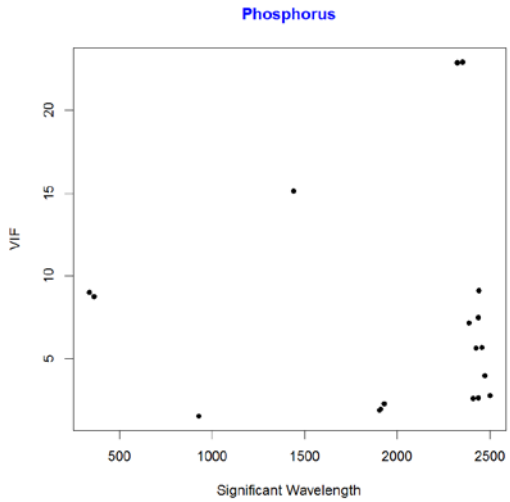


Figure 10.3: Scatterplot of VIF against Wavelength - Phosphorus

Significant Wavelength (Magnesium; nm): 349.3, 411.4, 695.3, 963.3, 1014, 1419.4, 1872.6, 1903.8, 1909.4, 1923.4, 1937.3, 1959.3, 1962.1, 1992.1, 2016.3, 2343.9, 2371, 2384.4, 2437, 2471.1,
2483.6, 2487.8, 2492, 2500.3

Variance Inflation Factors (VIF): 18.07, 14.55, 8.63, 8.85, 4.23, 13.19, 12.82, 1.96, 1.86, 2.39, 2.92, 6.65, 2.85, 15.54, 17.01, 25.15, 16.92, 8.32, 7.68, 4.14, 3.03, 2.2, 2.24, 2.76

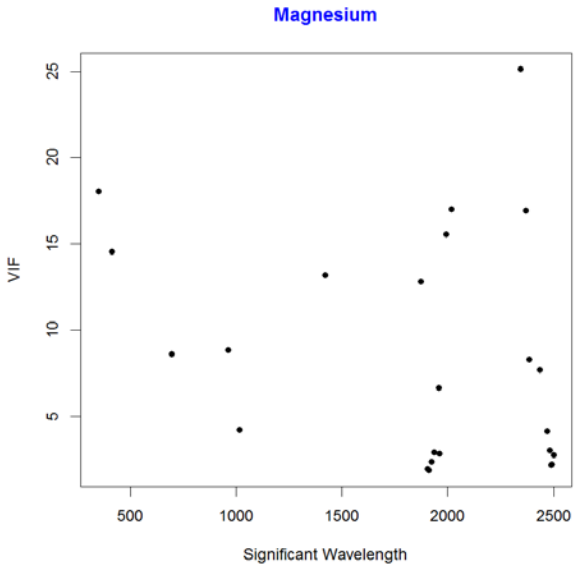


Figure 10.4: Scatterplot of VIF against Wavelength - Magnesium

Significant Wavelength (Zinc; nm): 337.3, 340.3, 516.5, 756.9, 963.3, 1113.1, 1415.7, 1837.1, 1897.3, 1903.8, 1926.2, 1951.1, 1962.1, 2002.9, 2323.2, 2357.5, 2382.2, 2386.7, 2410.9, 2437, 2462.6, 2471.1, 2485.7, 2492, 2496.1, 2500.3, 2508.5, 341.8

Variance Inflation Factors (VIF): 22.71, 24.03, 23.16, 6.45, 8.95, 5.38, 12.66, 41.32, 40.44, 2.16, 2.85, 4.4, 9.56, 7.05, 43.59, 25.92, 18.96, 10.62, 7.35, 7.45, 9.52, 5.57, 4.83, 3.01, 3.29, 2.98, 1.65, 12.6

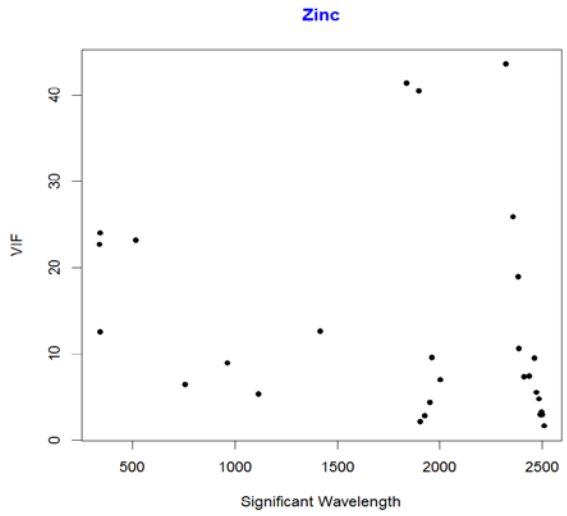


Figure 10.5: Scatterplot of VIF against Wavelength - Zinc

Significant Wavelength (Boron; nm): 337.3, 346.3, 457.9, 515.1, 656.4, 1400.7, 1869.1, 1903.8, 1906.6, 1942.8, 1945.6, 1948.3, 1953.8, 1997.5, 2410.9, 2430.5, 2452, 2456.3, 2473.2, 2481.6

Variance Inflation Factors (VIF): 11.82, 15.91, 10.24, 46.85, 57.18, 11.59, 10, 1.91, 2.34, 4.22, 13.99, 6.83, 6.61, 6.77, 6.61, 5.97, 9, 6.54, 4.81, 4.78

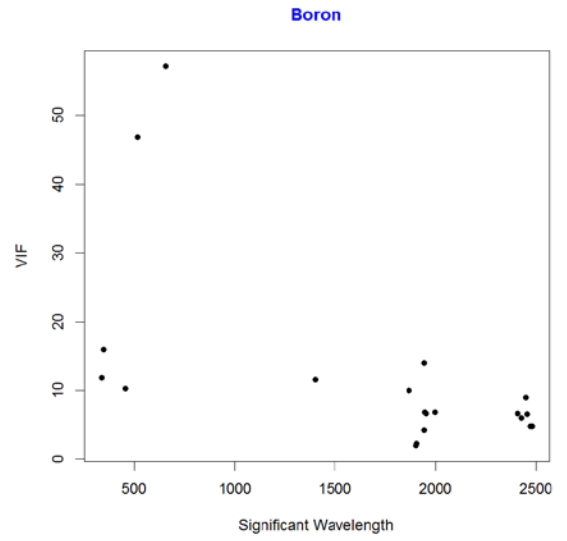


Figure 10.6: Scatterplot of VIF against Wavelength – Boron

10.3 Value of lambda.min and lambda.min.ratio as 0.003

The value of lambda.min and lambda.min.ratio has been selected as 0.003, to calculate the optimum value of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.68	0.73	0.67	0.74	0.79	0.65
Adj.R.Squared	0.62	0.65	0.61	0.68	0.73	0.57
Pred.R.Squared	0.52	0.41	0.55	0.61	0.58	0.48

Significant Wavelength (Nitrogen; nm): 337.3, 340.3, 571.3, 687.3, 758.2, 1438.2, 1872.6, 1906.6, 1912.2, 1928.9, 1942.8, 1956.6, 2355.2, 2368.8, 2386.7, 2393.3, 2419.7, 2452, 2483.6, 3 96.8,
1822.8

Variance Inflation Factors (VIF): 21.83, 24.35, 15.43, 47.68, 2.33, 26.94, 23.73, 2.29, 2.82, 2.53, 4.59, 7.01, 23.73, 4.43, 4.26, 16.51, 10.09, 6.78, 4.48, 17.19, 43.43

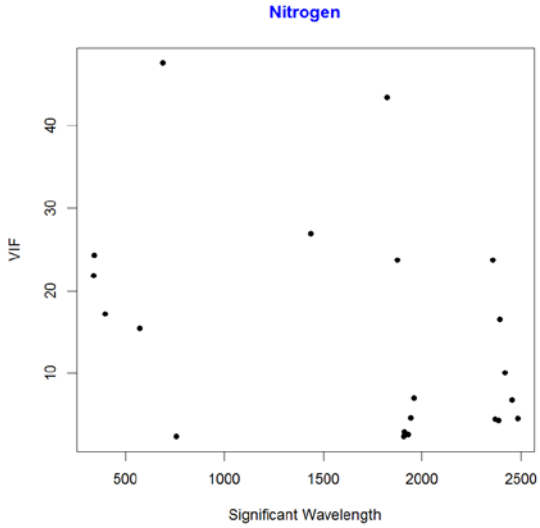


Figure 10.7: Scatterplot of VIF against Wavelength - Nitrogen

Figure 10.7 to Figure 10.12 displays the scatter plot of VIF against the wavelengths for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The concentration of significant wavelengths for the response variables for the grapevine dataset can be seen to have VIF around 10. Certain significant wavelengths have very high VIF; however, the median VIF of significant predictors (wavelength) has been tabulated in Table 10.2.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Median of VIF	Around 10	Around 18	Around 12	Around 10	Around 12	Around 11

Table 10.2: Median VIF of significant predictors for lambda.min of 0.003

Significant Wavelength (Potassium; nm): 334.3, 338.8, 341.8, 356.8, 398.2, 443.5, 1063.5, 1344.3, 1419.4, 1858.4, 1862, 1869.1, 1915, 1928.9, 1937.3, 1953.8, 1956.6, 1962.1, 1970.3, 1989.3, 2005.6, 2010.9, 2323.2, 2382.2, 2393.3, 2406.6, 2410.9, 2430.5, 2464.8, 2487.8, 2494.1, 2496.1, 2500.3

Variance Inflation Factors (VIF): 9.74, 26.68, 28.6, 47.19, 75.81, 26.82, 1.74, 60.06, 23.2, 59.81, 52.27, 51.07, 3, 3.8, 8.19, 10.59, 16.97, 17.34, 11.94, 17.81, 16.79, 13.13, 58.47, 21.56, 29.37, 16.41, 6.89, 7.99, 14.47, 3.38, 6.41, 4.3, 3.61

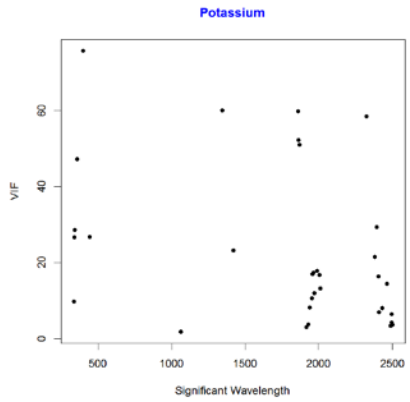


Figure 10.8: Scatterplot of VIF against Wavelength - Potassium

Significant Wavelength (Phosphorus; nm): 338.8, 691.3, 1438.2, 1822.8, 1897.3, 1900.8, 1909.4, 1920.6, 2323.2, 2355.2, 2362, 2382.2, 2386.7, 2426.2, 2437, 2458.4, 2500.3, 2473.2, 359.7, 925.6, 349.3

Variance Inflation Factors (VIF): 10.21, 13.06, 23.98, 41.64, 29.33, 21.15, 1.7, 2.72, 47.54, 29.85, 37.85, 12.19, 7.55, 5.03, 6.41, 5.69, 2.74, 4.86, 16.56, 2.01, 17.82

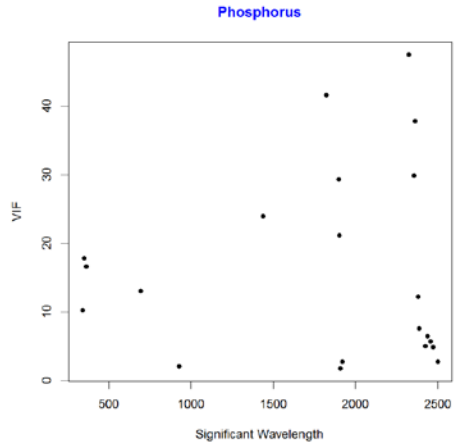


Figure 10.9: Scatterplot of VIF against Wavelength - Phosphorus

Significant Wavelength (Magnesium; nm): 349.3, 359.7, 512.2, 693.9, 963.3, 1021.6, 1419.4, 18 26.4, 1865.5, 1909.4, 1923.4, 1959.3, 1992.1, 2016.3, 2355.2, 2371, 2384.4, 2419.7, 2424, 2437, 2452, 2471.1, 2483.6, 2492, 2502.3, 2506.4, 2477.4

Variance Inflation Factors (VIF): 20.16, 17.95, 39.19, 25.18, 9.86, 4.62, 23.11, 53.43, 24.02, 1.9, 2.73, 11.43, 16.76, 22.79, 38.22, 33.25, 9.06, 15.19, 9.65, 9.82, 5.95, 5, 3.74, 2.69, 2.42, 2.09, 6.36

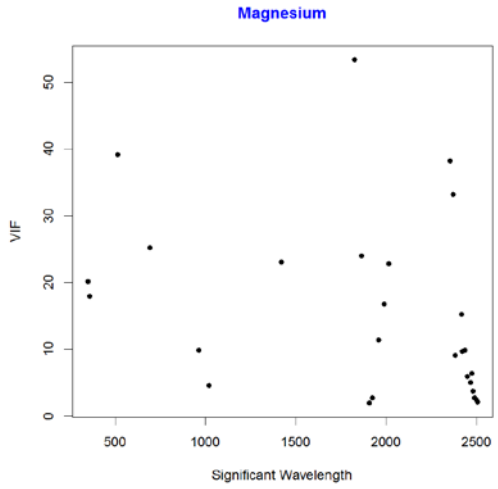


Figure 10.10: Scatterplot of VIF against Wavelength - Magnesium

Significant Wavelength (Zinc; nm): 337.3, 338.8, 340.3, 516.5, 646.9, 1113.1, 1415.7, 1512.8, 1830, 1837.1, 1897.3, 1903.8, 1951.1, 1962.1, 1992.1, 2323.2, 2357.5, 2377.7, 2382.2, 2386.7, 2 393.3, 2404.4, 2426.2, 2437, 2462.6, 2466.9, 2471.1, 2485.7, 2492, 2500.3, 2502.3, 2508.5

Variance Inflation Factors (VIF): 24.34, 9.75, 25.12, 68.19, 61.55, 2.46, 17.94, 56.1, 59.42, 55, 60.71, 2.25, 4.69, 9.32, 15.45, 58.69, 36.83, 8.33, 19.88, 15.11, 14.79, 15.61, 5.73, 11.6, 9.86, 11.18, 5.79, 6.3, 2.16, 4.83, 3.5, 1.88

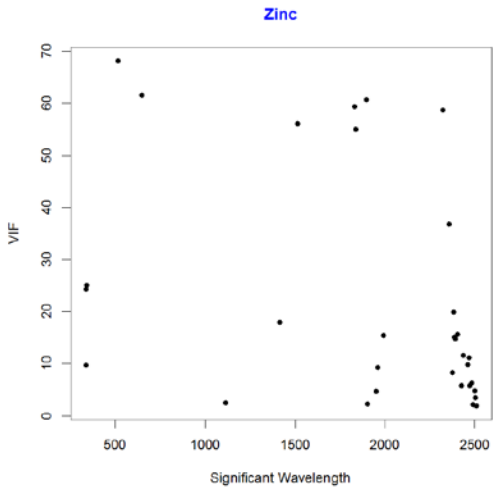


Figure 10.11: Scatterplot of VIF against Wavelength - Zinc

Significant Wavelength (Boron; nm): 337.3, 346.3, 449.3, 516.5, 656.4, 870.3, 1400.7, 1869.1, 1940, 1942.8, 1945.6, 1948.3, 1953.8, 1997.5, 2323.2, 2362, 2406.6, 2410.9, 2430.5, 2449.9, 2454.1, 2456.3, 2473.2, 2481.6, 2485.7, 2487.8, 2426.2

Variance Inflation Factors (VIF): 14.19, 17.34, 10.28, 44.38, 55.31, 2.05, 14.8, 12.21, 10.32, 5.66, 14.93, 7.56, 10.02, 19.05, 75.77, 52.75, 12.48, 9.31, 7.17, 6.76, 9.79, 8.26, 7.47, 7.51, 6.19, 3.11, 6.15

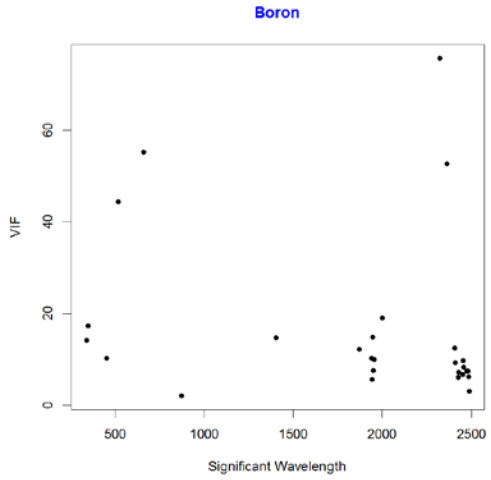


Figure 10.12: Scatterplot of VIF against Wavelength - Boron

10.4 Value of lambda.min and lambda.min.ratio as 0.0024

The value of lambda.min and lambda.min.ratio has been selected as 0.0024, to calculate the optimum value of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.60	0.73	0.75	0.61	0.79	0.61
Adj.R.Squared	0.54	0.65	0.68	0.55	0.73	0.52
Pred.R.Squared	0.37	0.41	0.59	0.35	0.59	0.29

Significant Wavelength (Nitrogen; nm): 392.3, 571.3, 685.9, 925.6, 1438.2, 1858.4, 1893.8, 1903.8, 1928.9, 1934.5, 1942.8, 1962.1, 1994.8, 2355.2, 2386.7, 2393.3, 2419.7, 2426.2

Variance Inflation Factors (VIF): 16.09, 9.35, 35.3, 1.84, 11.41, 34.78, 32.57, 1.64, 2.47, 2.34, 3.54, 5.8, 9.21, 29.62, 4.03, 12.77, 7.9, 3.32

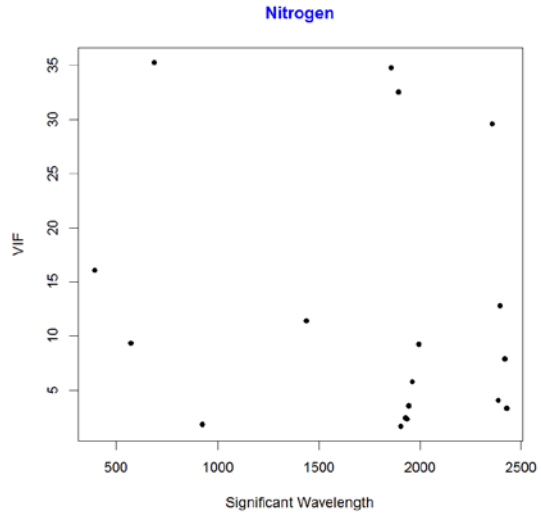


Figure 10.13: Scatterplot of VIF against Wavelength - Nitrogen

Figure 10.13 to Figure 8.18 displays the scatter plot of VIF against the wavelengths for nitrogen, potassium, phosphorus, magnesium, zinc and boron, respectively. The concentration of significant wavelengths for the response variables for the grapevine dataset can be seen to have VIF around 10. Certain significant wavelengths have high VIF; however, the median VIF of significant predictors (wavelength) has been tabulated in Table 10.3.

	Nitrogen	Potassium	Phosphorus	Magnesium	Zinc	Boron
Median of VIF	Around 8	Around 17	Around 8	Around 10	Around10	Around 15

Table 10.3: Median VIF of significant predictors for lambda.min of 0.0024

Significant Wavelength (Potassium): 334.3, 338.8, 341.8, 356.8, 398.2, 443.5, 1063.5, 1344.3, 1419.4, 1858.4, 1862, 1869.1, 1915, 1928.9, 1937.3, 1953.8, 1956.6, 1962.1, 1970.3, 1989.3, 2005.6, 2010.9, 2323.2, 2382.2, 2393.3, 2406.6, 2410.9, 2430.5, 2464.8, 2487.8, 2494.1, 2496.1, 2500.3

Variance Inflation Factor (VIF): 9.74, 26.68, 28.6, 47.19, 75.81, 26.82, 1.74, 60.06, 23.2, 59.81, 52.27, 51.07, 3, 3.8, 8.19, 10.59, 16.97, 17.34, 11.94, 17.81, 16.79, 13.13, 58.47, 21.56, 29.37, 16.41, 6.89, 7.99, 14.47, 3.38, 6.41, 4.3, 3.61

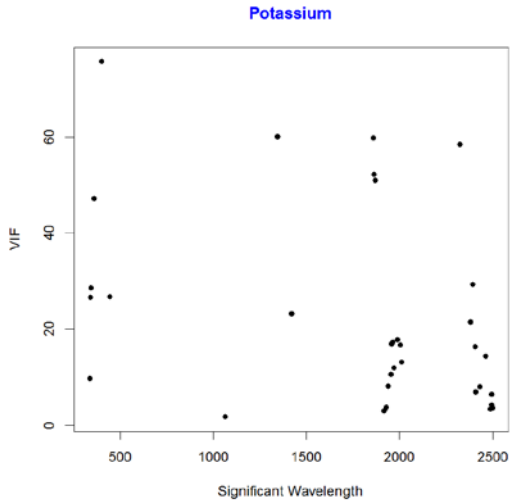


Figure 10.14: Scatterplot of VIF against Wavelength - Potassium

Significant Wavelength (Phosphorus): 334.3, 338.8, 340.3, 349.3, 359.7, 687.3, 925.6, 1438.2, 1826.4, 1897.3, 1903.8, 1906.6, 1909.4, 1912.2, 1942.8, 2323.2, 2355.2, 2362, 2371, 2382.2, 2386.7, 2410.9, 2426.2, 2437, 2439.1, 2462.6, 2473.2, 2483.6, 2500.3, 2502.3, 2508.5

Variance Inflation Factor (VIF): 20.88, 11.94, 26.47, 29.19, 20.89, 32.74, 2.33, 29.3, 65.99, 36.15, 2.76, 2.77, 2.54, 3.73, 6.22, 65.9, 38.64, 49.65, 30.75, 15.53, 10.54, 6.35, 6.36, 6.22, 3.84, 6.99, 5.5, 5, 5.1, 4.09, 1.87

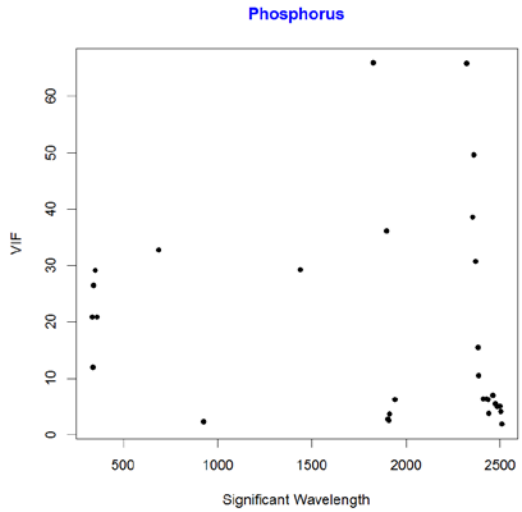


Figure 10.15: Scatterplot of VIF against Wavelength for Phosphorus

Significant Wavelength (Magnesium): 510.8, 693.9, 934.8, 1021.6, 1348.1, 1419.4, 1869.1, 1903.8, 1923.4, 2019, 2355.2, 2371, 2384.4, 2406.6, 2419.7, 2424, 2452, 2462.6, 2483.6, 2506.4
Variance Inflation Factor (VIF): 21.73, 19.89, 13.81, 12.62, 36.85, 9.91, 31.39, 1.72, 3.25, 22.39, 35.9, 18.57, 6.95, 8.84, 10.07, 6.74, 4.59, 2.87, 3.7, 2.22

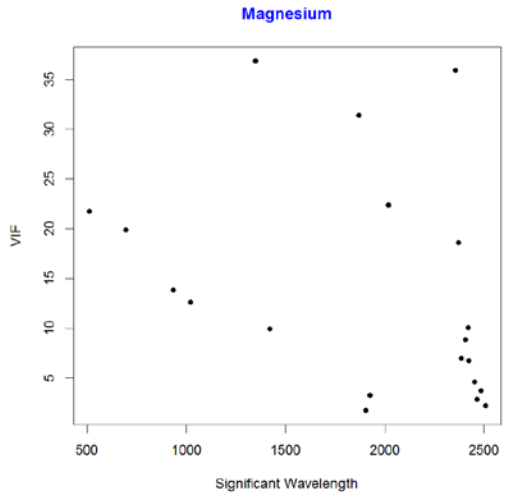


Figure 10.16: Scatterplot of VIF against Wavelength - Magnesium

Significant Wavelength (Zinc): 338.8, 341.8, 517.9, 646.9, 756.9, 963.3, 1113.1, 1415.7, 1512.8, 1830, 1837.1, 1883.2, 1897.3, 1903.8, 1926.2, 1951.1, 1967.6, 1992.1, 2323.2, 2357.5, 2382.2, 2 402.1, 2410.9, 2458.4, 2462.6, 2471.1, 2485.7, 2500.3, 2502.3, 2508.5

Variance Inflation Factor (VIF): 23.12, 31.6, 64.36, 57.38, 7.96, 10.52, 5.29, 14.85, 52.08, 71.01, 56.3, 69.55, 60.25, 2.33, 2.72, 3.73, 7.57, 15.52, 50.06, 35.39, 11.78, 9.95, 4.78, 6.1, 5.43, 5.18, 5.63, 4.4, 3.16, 1.79

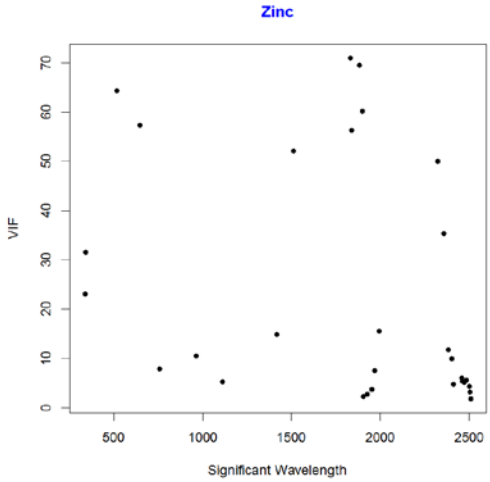


Figure 10.17: Scatterplot of VIF against Wavelength - Zinc

Significant Wavelength (Boron): 341.8, 343.3, 359.7, 517.9, 677.9, 725.7, 870.3, 1400.7, 1453.1, 1869.1, 1883.2, 1915, 1931.7, 1945.6, 1948.3, 1953.8, 2380, 2386.7, 2397.7, 2406.6, 2430.5, 2434.9, 2441.3, 2443.4, 2452, 2487.8, 2498.2, 2477.4

Variance Inflation Factor (VIF): 17.5, 14.77, 18.72, 29, 22, 44.74, 2.34, 33.67, 41.78, 30.01, 46.33, 2.4, 3.42, 9.69, 7.43, 8, 19, 8.74, 17.81, 12.02, 16.09, 9.33, 13.76, 19.76, 11.37, 2.98, 2.91, 8.03

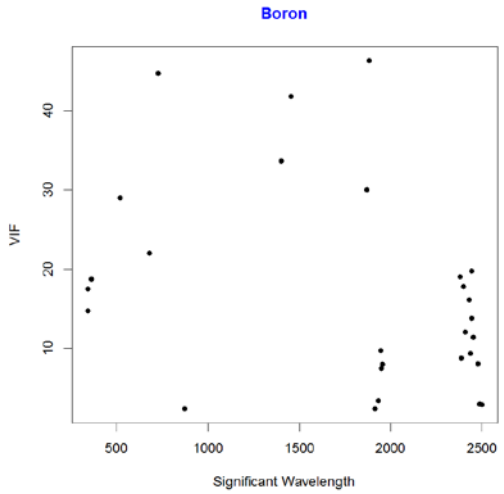


Figure 10.18: Scatterplot of VIF against Wavelength - Boron

The values of R-squared, adjusted R-squared, and predicted R-squared depends on the number of statistically significant predictors. In the case of nitrogen, 29, 21 and 18 significant predictors were selected when the values of lambda.min and lambda.min.ratio was 0.004, 0.003 and 0.0024, respectively. A slight reduction in the value of lambda min reduces the number of significant predictors. In the case of potassium, 20, 33 and 33 predictors were selected when the values of lambda.min and lambda.min.ratio was 0.004, 0.003 and 0.0024, respectively. A slight reduction in the value of lambda min increases the number of significant predictors, and then it flattens out. In the case of phosphorus, 18, 21 and 31 predictors are selected when the value of lambda.min and lambda.min.ratio was 0.004, 0.003 and 0.0024, respectively. A slight reduction in the value of lambda min increases the number of significant predictors. In the case of Magnesium 24, 27 and 20 predictors are selected when the value of lambda.min and lambda.min.ratio was 0.004, 0.003 and 0.0024, respectively. A slight reduction in the value of lambda min increases the number of significant predictors to the crest, and after that, it drops. In the case of zinc, 28, 32 and 30 predictors are selected when the value of lambda.min and lambda.min.ratio was 0.004, 0.003 and 0.0024, respectively. A slight reduction in the value of lambda min increases the number of significant predictors, and then it drops. In the case of boron, 20, 27 and 28 predictors are selected when the value of lambda.min and lambda.min.ratio was 0.004, 0.003 and 0.0024, respectively. A slight reduction in the value of lambda min increases the number of significant predictors. Except for the case of boron, when the number of predictors surges from 27 to 28, the value of R-squared, and adjusted R-squared increase with a rise in the number of significant predictors. Predicted R-squared also generally follows this trend.

It can be noticed that the increase or decrease in predictors is within the group selected by the elastic net. Additional predictors are usually the adjoining variable, and at times, the adjoining predictors replace the original predictor due to the nature of the elastic net. The change in the number of significant predictors, with a slight reduction in the value of lambda min and lambda min ratio, does not follow any one pattern. The decrease in the value of lambda min decreases the number of significant predictors for nitrogen and increases for phosphorus. In the case of potassium, a reduction in the value of lambda min initially enhances the number of significant predictors and then flattens out, whereas for magnesium, zinc, and boron the number of significant predictors initially increases, thereafter it drops. Hence, it is possible to get higher values of R-squared, adjusted R-squared, and predicted R-squared when the nutrients are studied separately.

The best overall values of R-squared, adjusted R-squared, and predicted R-squared were obtained by using the values of lambda.min and lambda.min.ratio of 0.003. Hence, we will use this value for the rest of the study.

Chapter 11

Comparison among Grapevine Datasets

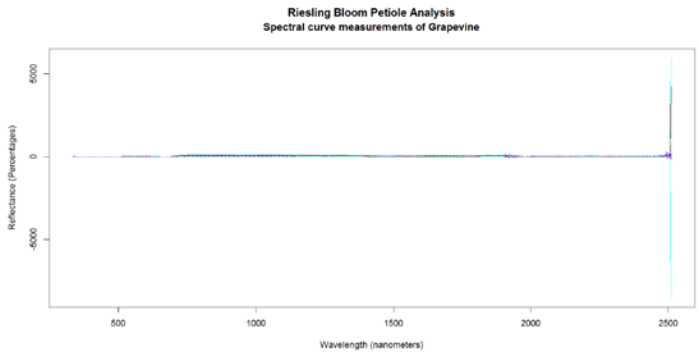
11.1 Introduction

In this chapter, a comparative study of four grapevine datasets is carried out for the spectral reflectance of leaf and associated petiole chemical analysis collected during the bloom and veraison periods for the two varieties namely Riesling and Cabernet Franc. For a better understanding, data have been taken from all the three angle of view namely; directly over individual grape leaves, the vine canopy at nadir and 15° off-nadir. The data, its source, and the data collection efforts are described in G. W. Anderson (2016) and Anderson et al. (2016). The first grapevine data are of petiole chemical analysis of the Riesling variety taken during the period of growth of bloom from the view angle directly over the individual grape leaves. The second grapevine dataset is based on Petiole analysis of the Riesling variety, taken during the veraison period of growth from the view angle directly at the nadir of the vine canopy. The third grapevine data are of Leaf of the Cabernet Franc variety but taken during the period of growth of bloom from the view angle at 15° off-nadir of the vine canopy. The fourth grapevine data are again of the leaf of the Cabernet Franc variety but taken during the period of growth of bloom from the view angle directly over the individual grape leaves.

We have noticed that the best result for the values for R-squared adjusted R-squared and predicted R-squared are obtained using the elastic net regularization path for fitting the generalized linear regression paths, by maximizing the appropriately penalized log-likelihood in the package `glmnet`. Hence, in this chapter, generalized linear model via penalized maximum likelihood is being used for the comparative study of four grapevine datasets. Also, the same parameters of seed= 5223, and `alpha`= 0.92 were selected, whereas the value of lambda.min and lambda.min.ratio were changed to calculate the optimum values of R-squared, adjusted R-squared, and predicted R-squared.

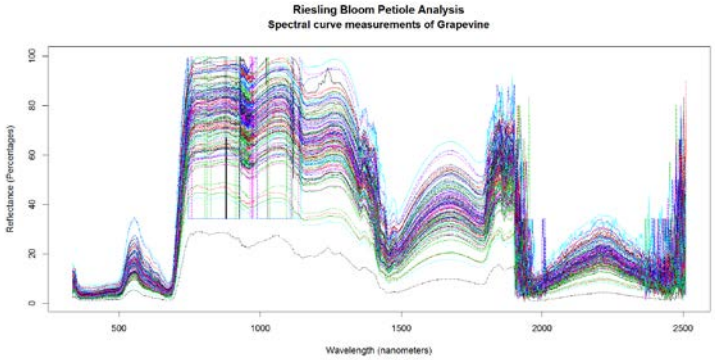
11.2 Exploratory Data of Riesling Bloom Petiole Chemistry Analysis and Leaf Reflectance

Since the radiance reflected from the leaf is expressed as a percentage of incident radiance through the range of wavelengths, it will have a value between 0 and 100. Hence, a spectral reflectance of less than 0 and more than 100 is considered a wrong observation. Detailed study shows that there are 72 and 828 observations with values less than zero and more than 100 respectively, indicating an error in data collection or entry. Thus, the grapevine dataset of petiole chemical analysis of the Riesling taken during bloom period directly over the individual leaves has 900 bad observations out of 141,984. It ranges from -8499 to less than 0 and more than 100 to 5962, as seen in the `figure 11.1`, below.



`Figure 11.1`: Spectral Curve measurement of Riesling Bloom Petiole Analysis dataset

Replacing these wrong observations with the mean value of the input (predictors) matrix, we get the spectral curve as given below.



`Figure 11.2`: Spectral Curve measurement of Riesling Bloom Petiole Analysis dataset without wrong observations

From the `Figure 11.2`, we can see that there is strong multicollinearity. The elastic net is known to select groups of correlated variables, which does not affect the predictability of the model. Hence, based on the best values of R-squared adjusted R-squared and predicted R-squared, we select 100 as the limit upper limit of the VIF. The process mentioned above will ensure that most of the variable will be within VIF of 10, but for a few variables, the VIF will be high. The wavelength, which is statistically significant for each nutrient and the associated VIF are given below.

Significant Wavelength (Nitrogen; nm): 337.3, 340.3, 571.3, 687.3, 758.2, 1438.2, 1872.6, 1906.6, 1912.2, 1928.9, 1942.8, 1956.6, 2355.2, 2368.8, 2386.7, 2393.3, 2419.7, 2452, 2483.6, 3 96.8,
1822.8
Variance Inflation Factors (VIF): 21.83, 24.35, 15.43, 47.68, 2.33, 26.94, 23.73, 2.29, 2.82, 2.53, 4.59, 7.01, 23.73, 4.43, 4.26, 16.51, 10.09, 6.78, 4.48, 17.19, 43.43

Significant Wavelength (Potassium; nm): 334.3, 338.8, 341.8, 356.8, 398.2, 443.5, 1063.5, 1344.3, 1419.4, 1858.4, 1862, 1869.1, 1915, 1928.9, 1937.3, 1953.8, 1956.6, 1962.1, 1970.3, 198 9.3,
2005.6, 2010.9, 2323.2, 2382.2, 2393.3, 2406.6, 2410.9, 2430.5, 2464.8, 2487.8, 2494.1, 2496.1, 2500.3
Variance Inflation Factors (VIF): 9.74, 26.68, 28.6, 47.19, 75.81, 26.82, 1.74, 60.06, 23.2, 59.81, 52.27, 51.07, 3, 3.8, 8.19, 10.59, 16.97, 17.34, 11.94, 17.81, 16.79, 13.13, 58.47, 21.56, 29.37, 16.41, 6.89, 7.99, 14.47, 3.38, 6.41, 4.3, 3.61

Significant Wavelength (Phosphorus; nm): 338.8, 691.3, 1438.2, 1822.8, 1897.3, 1900.8, 1909.4, 1920.6, 2323.2, 2355.2, 2362, 2382.2, 2386.7, 2426.2, 2437, 2458.4, 2500.3, 2473.2, 359.7, 925.6, 349.3
Variance Inflation Factors (VIF): 10.21, 13.06, 23.98, 41.64, 29.33, 21.15, 1.7, 2.72, 47.54, 29.85, 37.85, 12.19, 7.55, 5.03, 6.41, 5.69, 2.74, 4.86, 16.56, 2.01, 17.82

Significant Wavelength (Magnesium; nm): 349.3, 359.7, 512.2, 693.9, 963.3, 1021.6, 1419.4, 1826.4, 1865.5, 1909.4, 1923.4, 1959.3, 1992.1, 2016.3, 2355.2, 2371, 2384.4, 2419.7, 2424, 243 7, 2452, 2471.1, 2483.6, 2492, 2502.3, 2506.4, 2477.4

Variance Inflation Factors (VIF): 20.16, 17.95, 39.19, 25.18, 9.86, 4.62, 23.11, 53.43, 24.02, 1.9, 2.73, 11.43, 16.76, 22.79, 38.22, 33.25, 9.06, 15.19, 9.65, 9.82, 5.95, 5, 3.74, 2.69, 2.42, 2.0, 6.36

Significant Wavelength (Zinc; nm): 337.3, 338.8, 340.3, 516.5, 646.9, 1113.1, 1415.7, 1512.8, 1830, 1837.1, 1897.3, 1903.8, 1951.1, 1962.1, 1992.1, 2323.2, 2357.5, 2377.7, 2382.2, 2386.7, 2 393.3, 2404.4, 2426.2, 2437, 2462.6, 2466.9, 2471.1, 2485.7, 2492, 2500.3, 2502.3, 2508.5
Variance Inflation Factors (VIF): 24.34, 9.75, 25.12, 68.19, 61.55, 2.46, 17.94, 56.1, 59.42, 55, 60.71, 2.25, 4.69, 9.32, 15.45, 58.69, 36.83, 8.33, 19.88, 15.11, 14.79, 15.61, 5.73, 11.6, 9.86, 11.18, 5.79, 6.3, 2.16, 4.83, 3.5, 1.88

Significant Wavelength (Boron; nm): 337.3, 346.3, 449.3, 516.5, 656.4, 870.3, 1400.7, 1869.1, 1940, 1942.8, 1945.6, 1948.3, 1953.8, 1997.5, 2323.2, 2362, 2406.6, 2410.9, 2430.5, 2449.9, 24 54.1, 2456.3, 2473.2, 2481.6, 2485.7, 2487.8, 2426.2
Variance Inflation Factors (VIF): 14.19, 17.34, 10.28, 44.38, 55.31, 2.05, 14.8, 12.21, 10.32, 5.66 , 14.93, 7.56, 10.02, 19.05, 75.77, 52.75, 12.48, 9.31, 7.17, 6.76, 9.79, 8.26, 7.47, 7.51, 6.19, 3.11 , 6.15

11.3 Exploratory Data of Riesling Veraison Petiole Chemical Analysis at Nadir

The grapevine dataset of Riesling petiole chemical analysis taken during the veraison from directly at the nadir of the vine canopy has 1784 bad observations out of 141,984. Detailed study shows that there are 405 and 1379 observations with values less than zero and more than 100 respectively, indicating an error in data collection or entry. It ranges from -14905 to less than 0 and more than 100 to 10433.5 as seen in the `figure, 11.3`, below.

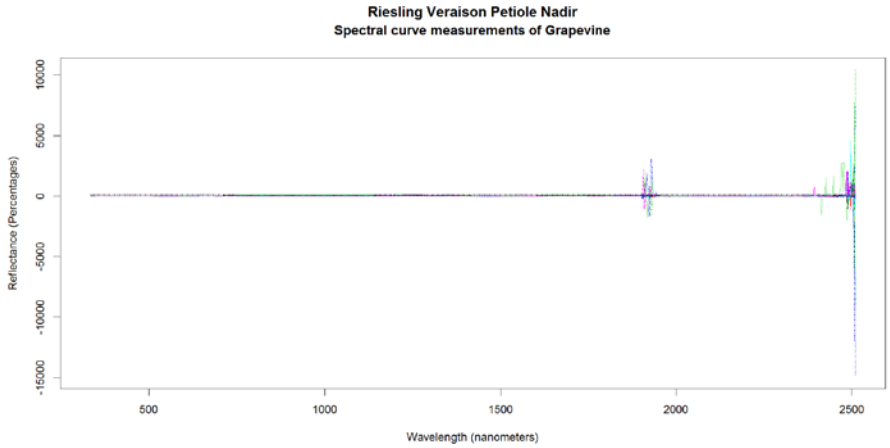


Figure 11.3: Spectral Curve measurement of Riesling Bloom at Nadir dataset

Replacing these incorrect observations with the mean value of the input (predictors) matrix, we get the spectral curves as given below.

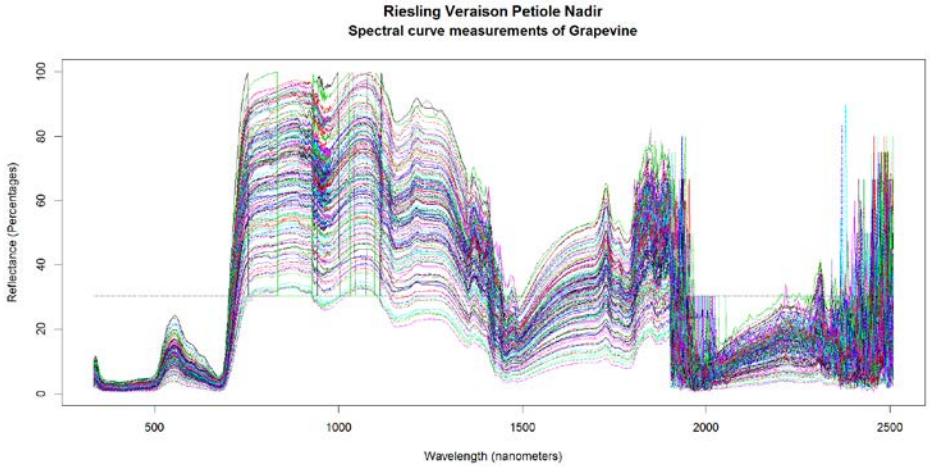


Figure 11.4: Spectral Curve of Riesling Bloom at Nadir dataset without wrong observations

From the Figure 11.4, we can see that there is a strong multicollinearity. Since the elastic net is known to select groups of correlated variables, which does not affect the predictability of the model. Hence, based on the best values of R-squared adjusted R-squared and predicted R-squared, we select 100 as the limit upper limit of the VIF. The method mentioned above will ensure that most of the variable will be within VIF of 10, but for a few variable, the VIF will be high. The wavelength which is statistically significant for each nutrient and the associated VIF is given below.

Significant Wavelength (Nitrogen; nm): 343.3, 1415.7, 1858.4, 1923.4, 1940, 1962.1, 1981.2, 2447.7, 2492
Variance Inflation Factors (VIF): 1.4, 6.36, 4.63, 2.98, 2.3, 1.69, 1.52, 2.63, 1.93

Significant Wavelength (Potassium; nm): 341.8, 692.6, 998.7, 1055.9, 1389.4, 1415.7, 1733, 1909.4, 1915, 1917.8, 1942.8, 1951.1, 1959.3, 1967.6, 1986.6, 2021.7, 2297.6, 2313.9, 2384.4, 2 413.1, 2419.7, 2426.2, 2437, 2481.6, 2496.1, 2498.2, 2506.4
Variance Inflation Factors (VIF): 18.57, 15.97, 62.45, 3.94, 88.83, 14.16, 81.79, 3.01, 3.6, 3.06, 3.77, 3.03, 3.27, 3.94, 2.2, 10.92, 20.53, 9.86, 6.29, 4.22, 4.52, 2.94, 5.36, 5.8, 4.06, 3.9, 3.28

Significant Wavelength (Phosphorus; nm): 530.6, 1423.2, 1920.6, 2334.7, 2437, 2460.5, 2500.3, 2504.4, 1862
Variance Inflation Factors (VIF): 3.97, 5.89, 1.88, 5.94, 2.03, 1.85, 2.01, 2.2, 6.19

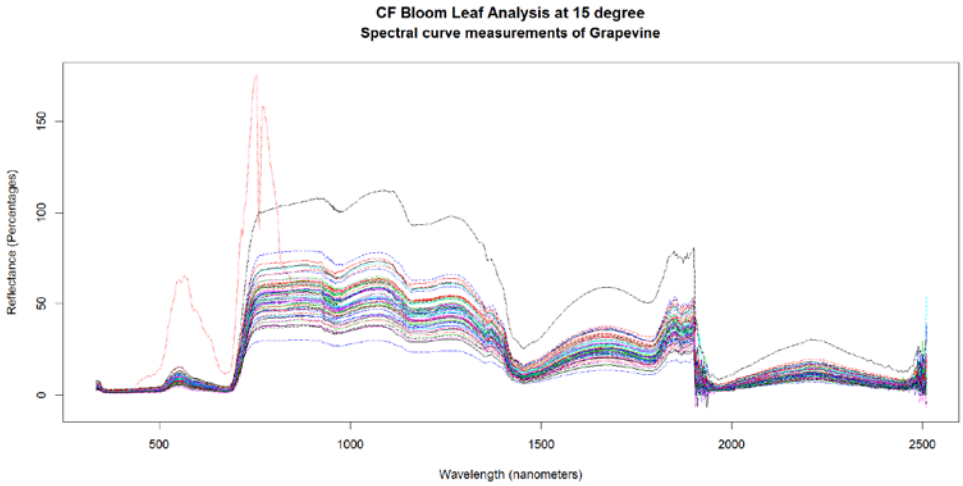
Significant Wavelength (Magnesium; nm): 1940, 2343.9, 2454.1
Variance Inflation Factors (VIF): 1.92, 1.07, 1.99

Significant Wavelength (Zinc; nm): 1415.7, 1886.7, 1897.3, 1912.2, 2377.7, 2434.9, 2452, 2454.1
Variance Inflation Factors (VIF): 7.01, 7.92, 9.45, 1.74, 1.8, 2.03, 2.77, 3.17

Significant Wavelength (Boron; nm): 335.8, 933.7, 2056.2
Variance Inflation Factors (VIF): 3.79, 1.55, 4.8

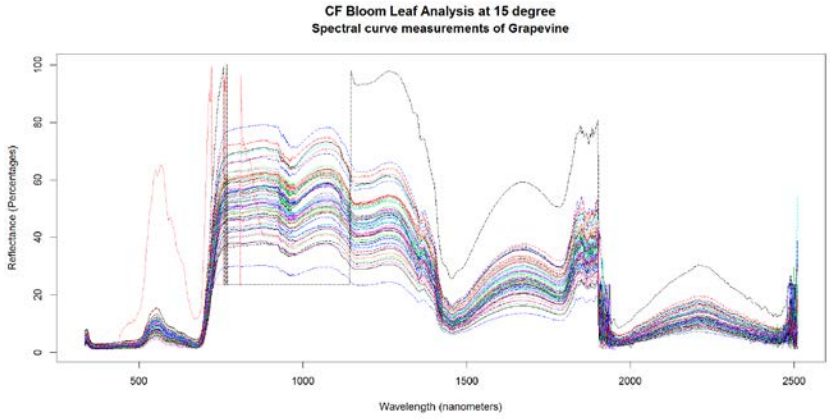
11.4 Exploratory Data of Cabernet Franc Leaf Analysis at 15^0

The grapevine data of the Leaf of the Cabernet Franc taken during blooming from a view angle at 15° off-nadir of the vine canopy has 303 bad observations out of 61,132. Detailed study shows that there are 14 and 289 observations with values less than zero and more than 100 respectively, indicating an error in data collection or entry. It ranges from -7.46 to less than 0 and more than 100 to 175.15 as seen in the `figure 11.5`, below.



`Figure 11.5`: Spectral Curve measurement of CF Bloom Leaf Analysis dataset

Replacing these wrong observations with the mean value of the input (predictors) matrix, we get the spectral curve, `figure 11.6` as given below.



`Figure 11.6:` Spectral Curve of CF Bloom Leaf datasets without wrong observations

From the `figure 11.6`, we can see that there is a strong multicollinearity. Since the elastic net is known to select groups of correlated variables, which does not affect the predictability of the model. Hence, based on the best values of R-squared adjusted R-squared and predicted R-squared, we select 68 as the limit upper limit of the VIF. The method mentioned above will ensure that most of the variables will be within VIF of 10, but for a few variables, the VIF will be very high. The wavelength which is statistically significant for each nutrient and the associated VIF, are given below.

Significant Wavelength (Nitrogen; nm): 343.3, 1981.2, 2458.4, 2483.6, 2489.9, 2494.1
Variance Inflation Factors (VIF): 3.69, 12.19, 6.99, 3.85, 1.89, 1.58

Significant Wavelength (Potassium; nm): 344.8, 1063.5, 1906.6, 1928.9, 1934.5, 2377.7, 2447.7, 2458.4, 2481.6
Variance Inflation Factors (VIF): 4.44, 1.22, 1.61, 2.64, 1.24, 21.73, 5.08, 15.82, 4.52

Significant Wavelength (Phosphorus; nm): 334.3, 343.3, 352.3, 426, 1351.8, 1903.8, 1912.2, 1926.2, 1931.7, 1937.3, 2424, 2445.6, 2449.9, 2473.2, 2475.3, 2487.8, 2500.3, 2504.4
Variance Inflation Factors (VIF): 13.65, 18.25, 26.21, 19.47, 8, 2.23, 2.85, 3.65, 6.11, 4.13, 23.07, 17.33, 6.95, 10.98, 13.78, 3.14, 3.33, 3.9

Significant Wavelength (Magnesium; nm): 337.3, 1837.1, 1906.6, 1923.4, 2489.9, 2496.1
Variance Inflation Factors (VIF): 2.71, 4.62, 1.88, 1.91, 1.79, 1.77

Significant Wavelength (Zinc; nm): 335.8, 337.3, 723, 827.7, 1815.7, 1906.6, 1915, 1920.6, 1926.2, 1931.7, 1934.5, 2443.4, 2447.7, 2466.9, 2489.9, 2498.2
Variance Inflation Factors (VIF): 10.96, 11.65, 31.1, 2.28, 26.39, 3.68, 3.29, 2.53, 2.96, 3.63, 2.51, 12.39, 9.44, 14.76, 2.23, 3.05

Significant Wavelength (Boron; nm): 335.8, 1906.6, 1934.5, 2447.7, 2481.6, 2492, 2496.1, 2498.2
Variance Inflation Factors (VIF): 3.58, 1.95, 1.67, 5.96, 4.3, 1.68, 2.53, 2.93

11.5 Exploratory Data of Cabernet Franc Leaf Analysis at Leaf

The grapevine data of the Leaf of the Cabernet Franc taken during blooming from directly over the individual grape leaves has seven incorrect observations out of 61,132. All the wrong observation has negative values, minimum being -25, as seen in the `figure 11.7`, below.

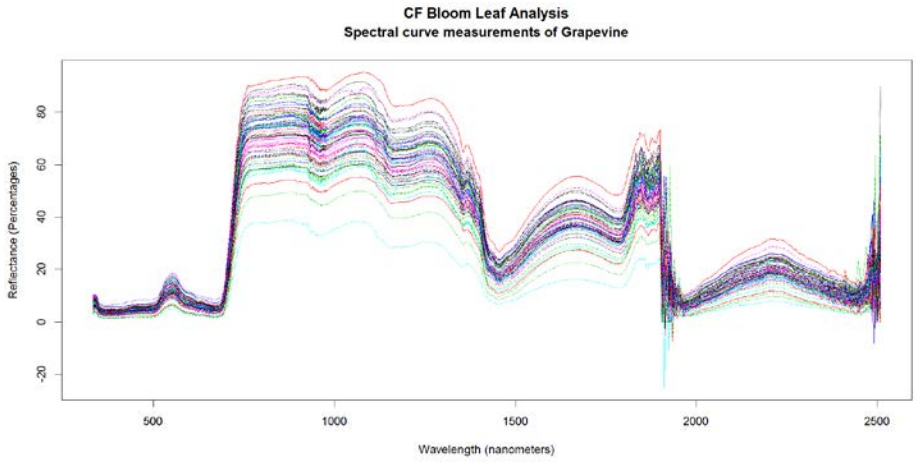


Figure 11.7: Spectral Curve measurement of CF Bloom Leaf Analysis dataset

Replacing these wrong observations with the mean value of the input (predictors) matrix, we get the spectral curve, figure 11.8 as given below.

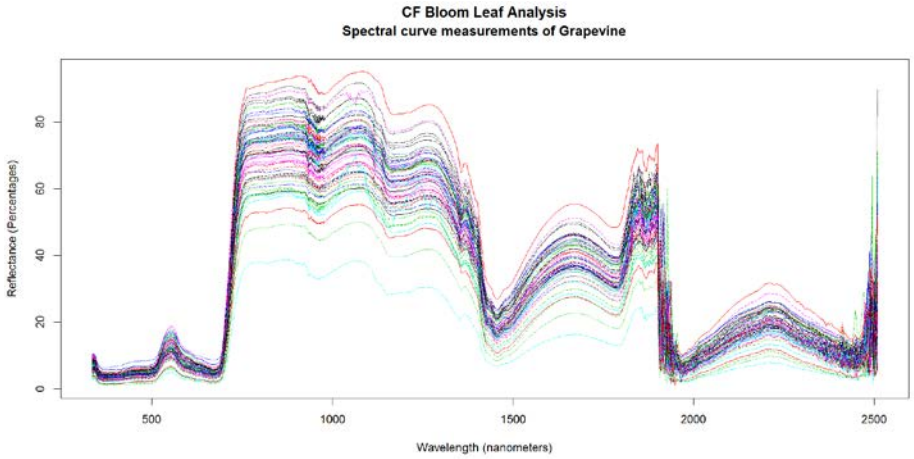


Figure 11.8: Spectral Curve measurement of CF Bloom Leaf Analysis dataset without wrong observation

From the figure 11.8, we can see that there is a strong multicollinearity. Since the elastic net is known to select groups of correlated variables, which does not affect the predictability of the model. Hence, based on the best values of R-squared adjusted R-squared and predicted R-squared, we select 68 as the limit upper limit of the VIF. The method mentioned above will ensure that most of the variables will be within VIF of 10, but for a few variables, the VIF will be high.

The wavelength, which is statistically significant for each nutrient and the associated VIF, are given below.

Significant Wavelength (Nitrogen; nm): 340.3, 359.7, 701.9, 1627.3, 1912.2, 1920.6, 1923.4, 1937.3, 1962.1, 2443.4, 2471.1, 2489.9, 2500.3, 2506.4
Variance Inflation Factors (VIF): 25.81, 13.48, 10.1, 8.24, 2.57, 2.26, 2.85, 3.52, 11.61, 7.87, 6.19, 2.67, 1.88, 3.26

Significant Wavelength (Potassium; nm): 1937.3, 2002.9, 2475.3, 2498.2, 2500.3, 2506.4
Variance Inflation Factors (VIF): 1.76, 1.78, 3.22, 1.27, 1.34, 2.64

Significant Wavelength (Phosphorus; nm): 708.5, 870.3, 1906.6, 1937.3, 2439.1, 2496.1, 2500.3
Variance Inflation Factors (VIF): 3.41, 3.79, 1.53, 1.68, 3.72, 1.53, 1.48

Significant Wavelength (Magnesium; nm): 1920.6, 1931.7, 1953.8, 2366.5, 2439.1, 2458.4, 2466.9, 2475.3, 2485.7, 2498.2
Variance Inflation Factors (VIF): 3.16, 1.74, 6.9, 10.69, 12.56, 15.77, 6.45, 6.56, 2.58, 1.98

Significant Wavelength (Zinc; nm): 1411.9, 1893.8, 1928.9, 1945.6, 2013.6, 2366.5, 2377.7, 2439.1, 2445.6, 2464.8, 2477.4, 2492, 2494.1
Variance Inflation Factors (VIF): 37.32, 24.99, 2.26, 4.97, 27.59, 23.96, 20.61, 14.51, 12.45, 5.84, 5.17, 3.66, 2.36

Significant Wavelength (Boron; nm): 692.6, 768.5, 1411.9, 1928.9, 1964.8, 2013.6, 2475.3
Variance Inflation Factors (VIF): 7.09, 13.31, 25.33, 1.9, 8.01, 17.18, 2.52

11.6 R-squared, adjusted R-squared and predicted R-squared values

We consider the grapevine dataset of petiole chemical analysis of Riesling, taken directly from the individual grape leaves during the bloom period. The value of seed and alpha has been chosen as 5223 and 0.92, respectively. The values of lambda.min and lambda.min.ratio were selected as 0.003, to calculate the optimum values of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.68	0.73	0.67	0.74	0.79	0.65
Adj.R.Squared	0.62	0.65	0.61	0.68	0.73	0.57
Pred.R.Squared	0.52	0.41	0.55	0.61	0.58	0.48

Now, we consider the grapevine dataset of petiole of Riesling, taken at the nadir of the grapevine canopy during the veraison period. The values of lambda.min and lambda.min.ratio were selected as 0.05, to calculate the optimum values of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.49	0.56	0.51	0.43	0.51	0.16
Adj.R.Squared	0.46	0.46	0.48	0.42	0.48	0.14
Pred.R.Squared	0.40	0.35	0.44	0.39	0.44	0.11

Next, we consider the grapevine dataset of Leaf analysis of the Cabernet Franc, taken at 15° off-nadir of the vine canopy during the bloom period. The values of lambda.min and lambda.min.ratio were selected as 0.011, to calculate the optimum values of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.35	0.65	0.82	0.59	0.79	0.58
Adj.R.Squared	0.28	0.59	0.75	0.55	0.71	0.52
Pred.R.Squared	0.20	0.46	0.51	0.43	0.63	0.45

Lastly, we consider the grapevine dataset of Leaf analysis of the Cabernet Franc, taken directly over the individual grape leaves during the bloom period. The values of lambda.min and lambda.min.ratio were selected as 0.0145, to calculate the optimum values of R-squared, adjusted R-squared and predicted R-squared.

	Nitrogen(%)	Potassium	Phosphorus	Magnesium	Zinc	Boron
R.Squared	0.73	0.59	0.54	0.72	0.80	0.52
Adj.R.Squared	0.65	0.54	0.48	0.66	0.75	0.45
Pred.R.Squared	0.45	0.46	0.40	0.57	0.69	0.35

11.7 Comparison of the four grapevine datasets

Notice that the prediction of the six nutrients is better when the readings of spectral reflectance taken directly from the grape leaves during the bloom period rather than taken at the nadir for the grapevine canopy during the veraison period. However, when the dataset of Riesling grape leaves is compared with the dataset of Leaf of the Cabernet Franc, taken at 15° off-nadir during the bloom period, we achieve a mixed result. The prediction of nitrogen, Phosphorus, magnesium, and boron is better by the leaf-level dataset, whereas potassium and zinc can be predicted better by the nadir leaf dataset. Similarly, we can compare the dataset of petiole chemical analysis of Riesling, taken at the nadir of the grapevine canopy during the veraison period with a leaf of the Cabernet Franc, taken at 15° off-nadir of the grapevine canopy during the bloom period. We again get a mixed result. Except for nitrogen, the remaining five nutrients can be predicted better by the dataset of Cabernet Franc, taken at 15° off-nadir during the bloom period. Now, compare the datasets of leaf analysis of the Cabernet Franc, taken at 15° off-nadir of the vine canopy with the one taken directly from the individual grape leaves during the bloom period. We again achieve mixed results. The predicted value of nitrogen, magnesium, and zinc is better for the readings taken directly from the individual grape leaves, whereas for phosphorus and boron can be predicted better by taking reading 15° off-nadir. The prediction of potassium is same for both the datasets.

Next, we compare the dataset of petiole chemical analysis of Riesling, taken at the nadir of the grapevine canopy during the veraison period and Leaf of the Cabernet Franc, taken directly from the grape leaves during the bloom period. Except for Phosphorus the prediction of remaining five nutrients are better for the readings taken directly from the Cabernet Franc grape leaves during the bloom period. Lastly, compare the dataset of petiole chemical analysis of Riesling, and leaf analysis of the Cabernet Franc, taken directly from the individual grape leaves during the bloom period. Except for potassium and zinc, the prediction of the remaining four nutrients is better for petiole chemical analysis of Riesling.

11.8 Findings of the selected four grapevine datasets

For the prediction of nitrogen, the best result (52%) can be achieved when the reflectance taken directly from the individual Riesling grape leaves during the bloom period.

For the prediction of potassium, the best result (46%) can be achieved when the reflectance from the leaf of the Cabernet Franc, taken at 15° off-nadir of the vine canopy or over the individual Riesling grape leaves during the bloom period.

For the prediction of phosphorus, the best result (55%) can be achieved when the reflectance from Petiole of the Riesling, is taken directly from the individual Riesling grape leaves during the bloom period.

For the prediction of magnesium, the best result (61%) can be achieved when the reflectance from Petiole of Riesling, is taken directly from the individual Riesling grape leaves during the bloom period.

For the prediction of zinc, the best result (69%) can be achieved when the reflectance from leaf analysis of the Cabernet Franc, is taken directly from the individual Riesling grape leaves during the bloom period.

For the prediction of boron, the best result (48%) can be achieved when the reflectance from petiole of Riesling, is taken directly from the individual Riesling grape leaves during the bloom period.

11.9 Recommendation based on analysis of four grapevine datasets

Based on the analysis of four grapevine datasets, it is recommended to take the spectral reflectance reading directly over the grapevine leaves during the bloom period to get the best-predicted values. Spectral reflectance of Riesling yields best-predicted values for nitrogen, phosphorus, magnesium, and boron while for potassium and zinc Cabernet Franc variety yields best-predicted values.

Chapter 12

Conclusion

Meeting the growing demand for wine over next couple of decades has generated much interest in the study of various characteristics of grapes, like fruit ripening rate, water status, infestation, and disease. To estimate the nutritional deficiencies of grapes, viticulturists are interested in six key nutrients: nitrogen, potassium, phosphorous, magnesium, zinc, and boron. The leaf reflectance of grapevine was collected from three different angles of view for Riesling and Cabernet Franc varieties during the bloom and the veraison period to predict the nutrients mentioned above. The nutrient analysis was performed at the petiole-level. Four datasets were selected to provide a correct representation of grape variety, growth period, the angle of view and parts of grapevine. The data, its source, and the data collections efforts are described in G. W. Anderson (2016) and Anderson et al. (2016).

The spectral reflectance of leaves was taken through wavelengths ranging from 330 to 2510 nanometers, at an interval of 1.5 to 2.7 nm. The reading for data collection was taken at 986 different wavelengths. The dataset of the Riesling variety had 144 observations whereas Cabernet Franc had 62 observations against 986 predictor variables. These high dimensional datasets, with a larger number of variables than the sample size, suffered from the curse of dimensionality and hence required shrinkage and variable selection.

Initially, these datasets were explored for missing values, wrong observations (outliers) and multicollinearity. There were no missing values. Since the radiance reflected from a leaf is expressed as a percentage of incident radiance through the range of wavelengths, it should have a value between 0 and 100. However, three grapevine datasets have spectral reflectance less than zero and an equal number of datasets with more than 100. Hence, all the four datasets had some bad observations; however, the severity of outliers was more for the datasets of Riesling than the Cabernet Franc variety. Robust regression and replacement of bad observations with the mean of the input matrix were examined to overcome the problem mentioned above. Based on their predictive ability, the second approach was selected for further study.

Since the spectral reflectance of leaves was collected through the range of wavelengths from 330 to 2510 nanometers, the datasets suffered from severe multicollinearity to the tune of 98% in certain cases. The Variance Inflation Factor (VIF) was restricted within ten as far as possible by utilizing the properties of Elastic Net and eliminating highly correlated predictors, wherever applicable.

Since these grapevine datasets are high dimensional, with multicollinearity, statistical inference is possible only by dimensionality reduction through sparse representation. The dimensional reduction will not only decrease the computational burden but also improve the estimation accuracy. For variable selection by sparsity, the coefficient of many predictors are reduced to zero, and non-zero components are considered as relevant variables. Thus, the estimation accuracy was improved by effectively identifying the subset of relevant predictors and the model interpretability enhanced with parsimonious representation. Four different methods were explored for variable selection, based on best-predicted values for the six nutrients utilizing the dataset of leaf spectral reflectance for Riesling grapes, taken directly from the leaves during the bloom period. The first three models dealt with linear regression while the fourth one was Functional Data Analysis.

The first regression model was based on convex penalized (pseudo-) likelihood using Elastic-Net regularization path via coordinate descent, which concurrently uses a mixture of the ℓ_1 (lasso) and ℓ_2 (ridge regression). This generalized linear model takes advantage of the property of elastic net, which simultaneously makes the automatic variable selection and continuous shrinkage, and selects groups of correlated variables using the R package, `glmnet`. Elastic net averages wavelengths that are highly correlated and then enters the averaged wavelength into the model. The predictive ability of this high dimensional grapevine dataset with high multicollinearity was good.

The second regression model was based on the regularization paths for Minimax Concave Penalty (MCP) with the so-called oracle property using the R package, `ncvreg`. This generalized linear model takes advantage of MCP, which takes off at the origin as the ℓ_1 penalty, but continuously relaxes that penalization until the rate of penalization drops to zero. However, the non-convexity nature of MCP introduces numerical challenges in fitting these models. For the

high-dimensional grapevine dataset, global convexity is neither possible nor relevant. Since the objective function of the grapevine dataset is convex in the local region that contains the sparse solutions, we still have stable estimates and smooth coefficient paths in the parameter space of interest. Though MCP tends to be more accurate as p increases, (possibly due to multicollinearity) the sparse solution of grapevine dataset selects a lesser number of nonzero coefficients than desired. This sparse solution adversely influences the predictive ability of the regression model based on Minimax Concave Penalty.

The third regression model was based on Iterative Sure Independence Screening (ISIS) using the R package, `SIS`. The sure screening method is based on correlation learning, which selects variables by filtering out the features that have a weak correlation with the response. This method ensures that all the relevant variables survive after the variable screening with a probability tending to one. SIS is based on the intuition that the predictors are independent; however, the absolute correlation coefficient between some of the predictors of high dimensional grapevine dataset are enormous. This collinearity between predictors of the grapevine dataset creates a problem in variable selection. It is possible that some unimportant predictors that are highly correlated with the significant predictors would be selected instead of important predictors that are relatively weakly related to the response. It is also possible that SIS would not have picked a significant predictor that was marginally uncorrelated but jointly correlated with the response variable. An iterative application of the SIS approach seeks to overcome the limitations of SIS, by making more use of the shared covariate information while retaining computational expediency and stability as in the original SIS. However, possibly due to multicollinearity, even ISIS selects fewer predictors than desired, which adversely affects the predictive ability of regression model.

Finally, functional data are defined as discrete observations of a phenomenon that can be represented by smooth curves, which reflect the dependence structure between neighboring points, so that the phenomenon can be evaluated at any point in time. The spectral reflectance of leaves was taken through wavelengths ranging from 330 to 2510 nanometers, at an interval of 1.5 to 2.7 nm. Hence, the spectral reflectance data measured along the continuum of wavelength can be represented by a smooth curve belonging to an infinite dimensional space. B-spline basis representation is used to compute the functional regression

between a functional explanatory variable (spectral reflectance of the grapevine data) $X(t)$ and the scalar response of the six nutrients. In spline smoothing, as in other smoothing methods, the mean squared error (MSE) is one way of capturing the quality of the estimate. For imposing smoothness on the estimated curve, MSE is reduced by sacrificing some bias to reduce sampling variance. Since the estimates are expected to vary gently from one value to another, we are effectively "borrowing information" from neighboring data values, thereby expressing our faith in the regularity of the underlying function x that we are trying to estimate. This pooling of information makes the estimated curve more stable, at the cost of some increase in bias (J. Ramsay & Silverman, 2005). Based on minimum mean MSE the number of basis function are chosen to calculate the predictive ability of functional data analysis. Since some basis functions (K) are not substantially smaller than the number of observations (n) of 144, the regression approach tends to overfit the data. Only a few basis functions are statistically significant. Hence the predictive ability of grapevine dataset is low.

The regression model, based on convex penalized (pseudo-) likelihood using Elastic-Net regularization path, provides the best predictive ability for the high-dimensional grapevine dataset with high multicollinearity.

The grapevine dataset is multivariate with correlation, which follows a different pattern. In other words, change in the parameters has a different impact on the predictability of the various nutrients. Hence, depending on the requirement to predict a particular nutrient, the parameters could be changed to obtain the best predictive value for that nutrient.

The comparison of four grapevine datasets was made based on the spectral reflectance of leaves of Riesling and Cabernet Franc grapes collected during the bloom and veraison period. It was noticed that different grapevine datasets are required for the best predictive value of the various nutrients. However, based on the analysis of datasets, the reading of the spectral reflectance for the Cabernet Franc or Riesling was taken at 15° off-nadir of the vine canopy or directly over the individual grapevine leaves during the bloom period, respectively, performed best.

It was found that all six nutrients in the four grapevine datasets have most of their significant predictors (wavelength) in three distinct ranges. The first range of wavelengths with significant predictors is from1820 to 2510 nanometers. The second range of wavelength with

significant predictors is between 330 and 450 nm. The third most prominent range of wavelengths with significant predictors is between 1340 and 1440 nm. Apart from these, all over the remaining range of wavelengths, there are a few isolated predictors, which are statistically significant.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Alma, Ö. G. (2011). Comparison of robust regression methods in linear regression. *Int. J. Contemp. Math. Sciences*, *6*(9), 409-421.

Anderson, G., van Aardt, J., Bajorski, P., & Heuvel, J. V. (2016). *Detection of wine grape nutrient levels using visible and near infrared 1nm spectral resolution remote sensing.* Paper presented at the SPIE Commercial+ Scientific Sensing and Imaging.

Anderson, G. W. (2016). An evaluation of the silicon spectral range for determination of the nutrient content of grape vines MS Thesis at the Rochester Institute of Technology.

Balabin, R. M., & Smirnov, S. V. (2011). Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. *Analytica chimica acta*, *692*(1), 63-72.

Bande, M. F., de la Fuente, M. O., Galeano, P., Nieto, A., Garcia-Portugues, E., & de la Fuente, M. M. O. (2016). Package 'fda. usc'.

Breheny, P., & Breheny, M. P. (2016). Package 'ncvreg.'

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, *5*(1), 232.

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (Wiley Series in Probability and Statistics).

Fan, J., Feng, Y., Saldana, D. F., Samworth, R., Wu, Y., & Feng, M. Y. (2016). Package 'SIS.'

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348-1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda. usc. *Journal of Statistical Software*, *51*(4), 1-28.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*: Springer Science & Business Media.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Friedman, J., Hastie, T., & Tibshirani, R. (2013). glmnet: Lasso and elastic-net regularized generalized linear models. Version1. In.

Geng, Z. (2014). *Variable Selection via Penalized Likelihood.* THE UNIVERSITY OF WISCONSIN-MADISON,

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Jacques, J., & Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, *8*(3), 231-255.

Levitin, D. J., Nuzzo, R. L., Vines, B. W., & Ramsay, J. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne, 48*(3), 135.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., . . . di Palma, M. A. (2016). Package 'robustbase.'

Maimon, O., & Rokach, L. (2005). Data Mining and Knowledge Discovery Handbook. Secaucus. In: NJ, USA: Springer-Verlag New York, Inc.

Mallows, C. L. (1973). Some comments on C p. *Technometrics, 15*(4), 661-675.

Matsui, H., Kawano, S., & Konishi, S. (2009). Regularized functional regression modeling for functional response and predictors. *Journal of Math-for-industry, 1*(3), 17-25.

. Mineral nutrients. (1998). In *The Encyclopedia of Ecology and Environmental Management, Blackwell Science*: Blackwell Publishers.

. Mineral Nutrition and Suppression of Plant Disease. (2014). In *Encyclopedia of Agriculture and Food Systems*: Elsevier Science & Technology.

Mizuta, M., & Kato, J. (2007). *Functional data analysis and its application.* Paper presented at the International Conference on Rough Sets and Knowledge Technology.

Ordóñez, C., Rodríguez-Pérez, J. R., Moreira, J. J., & Sanz, E. (2013). Using hyperspectral spectrometry and functional models to characterize vine-leaf composition. *IEEE Transactions on Geoscience and Remote Sensing, 51*(5), 2610-2618.

Ramsay, J., & Silverman, B. (2005). *Functional Data Analysis*: Springer Science & Business Media.

Ramsay, J. O., & Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 539-572.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*: Springer Science & Business Media.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 73-79.

Saldana, D. F., & Feng, Y. (2016). SIS: An R package for Sure Independence Screening in Ultrahigh Dimensional Statistical Models. *Journal of Statistical Software*.

Schulz-Streeck, T., Ogutu, J., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464.

Susanti, Y., & Pratiwi, H. (2014). M-estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics, 91*(3), 349-360.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Wu, P.-S., & Müller, H.-G. (2010). Functional embedding for the classification of gene expression profiles. *Bioinformatics, 26*(4), 509-517.

Yohai, V. J. (1987). High breakdown-point and high-efficiency robust estimates for regression. *The Annals of Statistics*, 642-656.

Zarco-Tejada, P. J., Berjón, A., López-Lozano, R., Miller, J., Martín, P., Cachorro, V., . . . De Frutos, A. (2005). Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sensing of Environment, 99*(3), 271-287.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894-942.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.