Variable Selection Methods for Model-based Clustering and Application to High-dimensional Data

BY

Jini Xu

A Thesis Submitted to the Faculty of Science

Department of Mathematics and Statistics

In Partial Fulfilment

of the Requirements

for the Degree of

Master of Science

February 13, 2022

Title : Variable Selection Methods for Model-based Clustering and Application to High-dimensional Data

Author : Jini Xu

M.Sc., (Mathematics and Statistics)

McMaster University, Hamilton, Canada

Supervisor: Dr. Sharon McNicholas

Co-Supervisor: Dr. Pratheepa Jeganathan

Abstract

Abstract − Clustering helps in understanding the natural grouping and internal structure of data. Model-based clustering considers each cluster as a component in a mixture model. As the data dimensionality and complexity increase, model-based clustering tends to over-parametrize results. Thus, it is important to select a subset of critical variables instead of using all the variables for clustering. This study considers two variable selection methods for model-based clustering on real world high-dimensional data; variable selection for clustering and classification (VSCC) and variable selection for model-based clustering (clustvarsel). For simplicity, Gaussian mixture models were applied. Three criteria are used to compare the clustering accuracy and efficiency, which are the adjusted rand index (ARI), mis-clustering error, and performance time (in seconds).

Contents

List of Tables

List of Figures

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance. I would first like to thank my supervisors, Dr. Sharon McNicholas and Dr. Pratheepa Jeganathan. Your insightful feedback pushed me to sharpen my thinking and helped me learned a lot through the process. Thank you very much for your support and understanding. I would also like to thank Dr. Paul McNicholas for giving me advice. In addition, I would like to thank my parents and friends for their care and support.

Chapter 1

Introduction

Generally, classification is the procedure that assigns a group of labels to unlabelled observations, and the group refers to a class or a cluster. Based on the prior information of labelled observations and the extent to which it is used, classification can be categorized into three types: supervised, semi-supervised, and unsupervised (McNicholas, 2016). Supervised and semi-supervised classification have some pre-labelled observations that can be used to infer the unlabelled observations. Unsupervised classification, also called clustering, has no prior labelled observations. For the definition of a cluster, Wolfe (1963) suggested two definitions of a cluster whereby data in the same group are more similar to each other than the data in the other groups. (McNicholas, 2016) pointed out the definition is problematic and it is more precise to use mixture component density-based definitions. These definitions define a cluster as a component within a finite mixture model, which is unimodal, and an appropriate mixture model should be used for the data.

As the data variety and complexity increases, labelling data can be time consuming and expensive, hence clustering becomes an ideal choice for saving time and cost. However, there are mathematical and statistical challenges for clustering high-dimensional data sets. Johnstone & Titterington (2009) discussed visualization problems that can arise when dealing with a large number of measurements. A problem with model-based clustering methods is over-parametrization (Bouveyron & Brunet-Saumard, 2014). Therefore, variable selection becomes important for clustering high-dimensional data, which can reduce the dimensionality of the data set, facilitate model fitting, ease the interpretation of the results, and increase the accuracy of clustering. Not only for the high-dimensional data sets, also for moderate or low dimensionality, implementing variable reduction techniques can be beneficial for clustering (Fowlkes et al., 1988).

This study is interested in finding efficient variable selection techniques for model-based clustering on high-dimensional data sets, in terms of the clustering accuracy and the performance time. Another goal is to find which technique is suitable for certain types of data sets. This study gives a review of the variable selection methods in model-based clustering and the relevant R packages.

Fop & Murphy (2018) summarized that there are two main assumptions for variable selection: the local independence assumption and the global independence assumption. Local independence assumes the relevant variables are conditionally independent within the groups and global independence assumes the irrelevant variables are independent of relevant clustering variables. Relevant variables are defined as variables that contain the primary clustering information, irrelevant

variables are the variables that do not provide beneficial information, and redundant variables provide information that is already provided by the relevant variables. The local independence assumption is a standard assumption of the latent class analysis model (Clogg, 1988). Gaussian mixture models use global independence assumptions, assuming components with diagonal covariance matrices (zero covariances between variates). The global independence assumption eases the modelling of the relation between relevant and irrelevant variables but it also restricts the capability for considering the existence of redundant variables. Most of the variable reduction methods use one of the assumptions. Liu & Motoda (2007) defined the variable selection methods as filter methods or wrapper methods based on how the variable selection algorithms interact with the model fitting processes. The filter methods select variables as a pre (or post) processing step, with the wrapper methods performing learning and variable selection steps at the same time. Filter methods are computationally efficient, while wrapper methods are more popular because of better results.

The variable selection methods for clustering multivariate continuous and categorical data sets have 3 main approaches: Bayesian, penalization, and model selection (Fop & Murphy, 2018). Fowlkes et al. (1988) proposed an algorithm to find the subset of variables, named forward selection, in the context of complete linkage hierarchical cluster analysis. In the context of variable selection for Gaussian mixture models, Lawrence et al. (2003) proposed one of the Bayesian approaches that selects the most informative principal components (or variables) of the data with component analysis (PCA), which is the first dimension reduction step. Law et

al. (2004) provided an alternative approach, also known as the "hard selection" method. It has a binary indicator, if the indicator value is 1 then the variable is considered as relevant, otherwise, the variable is irrelevant. Tadesse et al. (2005) has a similar idea for using a binary indicator for variable selection. Lots of research that has been done on penalization approaches. Pan & Shen (2007) used a penalty function on the model parameters and variable selection to reduce the sparsity in the estimates. After centering the data, it selects the variables by shrinking the small estimates of means towards zero. There is a similar approach proposed by Bhattacharya & McNicholas (2014), and more different penalty terms were created by researchers. Within the class of model selection approaches, Raftery & Dean (2006) proposed a greedy algorithm to select variables using Bayes factors. The method selects variables, the number of clusters, and the clustering model simultaneously. The methods can be performed in R using the package clustvarsel (Scrucca & Raftery, 2018). Future work was extended by Maugis et al. (2009a) that suggested the conditional distribution can be related only to a subset of the clustering variables that avoid the inclusion of some unneeded variables. Later on, Maugis et al. (2009b) considered an additional rule for adding or removing proposed variables. They assumed the set of added or removed variables can be independent of the set of current variables. Maugis-Rabusseau et al. (2012) extend the variable selection framework for dealing with missing data. For the above methods, the likelihood needs to be optimized multiple times. To deal with this computational issue, Marbac & Sedki (2017b) proposed a method that depends on the ICL criterion and does not require optimization of the EM algorithm lots

of times. There is an R package available for this method, named VarSelLCM (Marbac & Sedki, 2017a). Other variable selection techniques do not belong to one of the 3 approaches, Dy & Brodley (2004) proposed to embed in the EM a forward selection algorithm for the maximization of two simple alternative criteria for variable selection, scatter separability and maximum likelihood. A similar one is Andrews & McNicholas (2014), which introduced a variable selection method for both clustering and classification of high-dimensional data, known as VSCC. It is a hybrid filter-wrapper approach and it minimizes the within-group variance and maximizes the between-group variance. The R package is named as VSCC (Jeffrey L. Andrews, 2013). There are other available R packages, bclust (Nia & Davison, 2012), sparcl (Witten et al., 2013), and SelvarMix (Sedki et al., 2014).

To cluster multivariate categorical data sets, a latent class analysis model (LCA; Bartholomew et al., 2011) is commonly used. The mixture density is a mixture of multinomial distributions. The variable selection methods for latent class analysis have the same 3 main approaches: Bayesian, penalization, and model selection. Researchers have paid attention to Bayesian approaches in recent years, Law et al. (2004) proposed a concept of feature saliency and introduced an expectation-maximization (EM) algorithm to measure it's value. Silvestre et al. (2015) proposed an adaptation of this method to categorical data clustering and variable selection. They used the same concept of feature saliency but utilized a new mixture model instead of a Gaussian Mixture Model. For the penalization approaches, Houseman et al. (2006) proposed a penalization method for latent class regression analysis of high-dimensional genomic data. The response variables are

categorical and they are related to the numerical covariates. Either the ridge or LASSO penalty term can be used to obtain the estimates for the regression coefficients. Then, discard the variables with coefficients shrunken to zero. This method does not select variables directly during clustering, Wu (2013) proposed a related method focused on selecting variables involved in the clustering. For the model selection approaches, Dean & Raftery (2010) proposed a framework similar to Raftery & Dean (2006) that separated the data into a set of current clustering variables, a set of proposed added or removed variables, and a set of irrelevant variables. The method compares two models and decides whether to add or remove a variable from the proposed set of variables based on BIC values. This method encounters problems of multimodality of the LCA log-likelihood and has sensitivity of the initialization. Bartolucci et al. (2016) added an extra step to solve the problem of initializing the estimation algorithm with a large amount of random starts. Another problem with the method (Dean & Raftery, 2010) is the independence assumption between proposed variables and the clustering variable in one of the models, which does not consider the proposed variable could be redundant. Fop et al. (2017) proposed a method to solve this problem by assuming the set of proposed variables are irrelevant to the clustering and that they are dependent on the clustering variables. There are a lot of other model selection approaches that perform well on clustering, Fop & Murphy (2018) gave a review on lots of other methods. The following are some available R packages for latent class analysis variable selection, ClustMMDD (Toussile & Gassiat, 2009; Bontemps & Toussile, 2013), LCAvarsel (Dean & Raftery, 2010; Fop et al., 2017), and VarSelLCM (Mar-

bac & Sedki, 2017b; Marbac & Sedki, 2017a).

Variable selection is not only useful in clustering, it can also be applied to different areas, such as regression analysis. Greenland (1989) concluded some rules for variable selection and model selection for multivariate modelling on epidemiologic data. Kuo & Mallick (1998) showed a variable selection method based on posterior probability for generalized linear models. Bleich et al. (2014) applied variable selection for Bayesian additive regression trees and implemented it on gene regulatory network in yeast. Desboulets (2018) reviewed the different variable selection methods in regression analysis, the main categories of algorithms are: test-based, penalty-based, and screening-based.

The thesis contains the following five chapters: theory, methodology, application, discussion and conclusion. Chapter 2 includes related distributions, two methods that were implemented in this study and their R packages, variable selection for clustering and classification (VSCC) and variable selection for model-based clustering (clustvarsel). Chapter 3 introduces parameter estimation methods, variable selection methods, model selection criteria, and performance assessment of methods. Chapter 4 shows the results of clustering various real-world data sets by using VSCC and clustvarsel. The last two chapters contain discussions and conclusions based on the results.

Chapter 2

Theory

2.1 Finite Mixture Models

Let $\mathbf{X}$ be an n by p data matrix, and $\mathbf{x}$ be a p dimensional random vector with $\mathbf{x} = (\mathbf{x_1}, ..., \mathbf{x_p})$. For each $\mathbf{x} \in \mathbf{X}$, the finite mixture model has the following density formula (McNicholas, 2016)

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g), \tag{2.1}$$

where π_g is the fraction of the population in cluster g, $\pi_g > 0$, and $\sum_{g=1}^{G} \pi_g = 1$. For $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$, it is the gth component density, π and $\boldsymbol{\theta}$ are the parameter vectors for the cluster g, where $\boldsymbol{\vartheta} = (\pi_1, ..., \pi_G; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G)$. Also, $f(\mathbf{x} \mid \boldsymbol{\vartheta})$ is named as G-component finite mixture density. Usually, the component densities $f_1(x \mid \boldsymbol{\theta}_1), ..., f_G(x \mid \boldsymbol{\theta}_G)$ are treated as having the same distribution (McNicholas, 2016). After McLachlan & Basford (1988) published a monograph of the potential usefulness of mixture models for inference and clustering, researchers found lots of interesting fields to

apply mixture models to. There are books about mixture models written by Lindsay (1995), Peel & MacLahlan (2000), Frühwirth-Schnatter (2006), and Mengersen et al. (2011). These books include classification, machine learning, multivariate analysis, applications, and other areas.

2.2 Gaussian Mixture Models (GMM)

The Gaussian mixture model is a common approach for clustering multivariate continuous data. The probability density function of a Gaussian mixture model (GMM) is given as following

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.2}$$

where $\phi(\mathbf{x} \mid \mu_g, \Sigma_g)$ is the density function of a random variable $\mathbf{x}$. It has a multivariate Gaussian distribution (McNicholas, 2016),

$$\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^p \left|\boldsymbol{\Sigma}_g\right|}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)\}, \tag{2.3}$$

where $\boldsymbol{\mu}_g$ is an p by 1 vector with the mean of all variables in the gth cluster, and $\boldsymbol{\Sigma}_g$ is a p by p covariance matrix of the gth cluster. Model based-clustering via GMMs can be performed in R using the package mclust (Fraley et al., 2012, Scrucca et al., 2016).

2.3 Variable Selection Methods

There are 3 major variable selection methods for Gaussian mixture models: Bayesian approaches, penalization approaches, and model selection approaches (Fop & Murphy, 2018). The method clustvarsel belongs to the model selection approach, and the VSCC is a hybrid filter-wrapper approach based on the within-group variance.

2.3.1 Variable Selection for Clustering and Classification (VSCC)

Variable selection for clustering and classification minimizes the within-group variance and maximizes the between-group variance at the same time. Suppose there are n observations, p variables, and g clusters. Let z_{ig} be the indicator, if $z_{ig} = 1$ that means the observation belongs to cluster g; if $z_{ig} = 0$, it does not belong to cluster g. The following formula represents the the within-group variance for each variable $j = 1, 2, ..., p$,

$$V_j = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig}(x_{ij} - \mu_{gj})^2}{n}, \tag{2.4}$$

where x_{ij} represents the ith observation with variable j, and μ_{gj} represents the mean of variable j in group g (Andrews & McNicholas, 2014). The between-group variance of variable j is not explained by this formula. If the data is standardized such that each variable has equal variance, then minimizing the within-group variance for any variable is simultaneously maximizing the between-group variance

(Andrews & McNicholas, 2014).

To select the variables, firstly set a threshold, ρ. Let U be the space of current selected variables, select variable j if all $u \in U$, and

$$\left|\rho_{ju}\right| < 1 - V_j^m, \tag{2.5}$$

where ρ_{ju} represents the correlation between variable j and u, and m represents the degree of the variance-correlation relationship. Andrews & McNicholas (2014) visualized that there will be more variables included as the m value increases, thus they suggest to take integer m from 1 to 5. The following is the alogirthm for VSCC:

Step 1: Calculate the within-group variances, V_j.

Step 2: Sort V_j in an ascending order and denote it as $\mathbf{V}_s$.

Step 3: V_1 is automatically selected and added to selected variable set U since it minimizes V_s. Then, set a count $k = 2$.

Step 4: While $k < p$, if $|\rho_{ku}| < 1 - V_k$ for all $u \in U$, then add variable k to the space U.

VSCC uses model-based clustering to initialized the (hard) z_{ig} values, which is contained in the mclust package (Fraley & Raftery, 2006). It also later implements mclust on the five variable subsets (less or equal to five) provided by VSCC, and selects the best model by comparing the BIC values.

2.3.2 Variable Selection for Model-based Clustering (Clustvarsel)

Raftery & Dean (2006) proposed a greedy algorithm to select variables by using Bayes factors. The method separates data into 3 subsets: the set of current clustering variables (X^C), the variable proposed to be added or removed (X^P), and the set of irrelevant variables (X^I). The decision of adding or removing the proposed variable is taken by comparing the following two models:

$$M_1 : p(X \mid z) = p(X^C, X^P \mid z)p(X^I \mid X^C, X^P) \tag{2.6}$$

$$M_2 : p(X \mid z) = p(X^C \mid z)p(X^P \mid X^C)p(X^I \mid X^C, X^P) \tag{2.7}$$

In the model M_1, X^P provides additional information for clustering and the joint distribution $p(X^C, X^P \mid z)$ corresponds to a Gaussian mixture distribution. In model M_2, X^P does not provide additional information to the clustering and $p(X^P \mid X^C)$ corresponds to a linear regression. The model formulation does not require the irrelevant variables (X^I) to be independent of the clustering variables. The ratio is defined as the Bayes factor (Kass & Raftery, 1995):

$$B_{12} = \frac{p(X \mid M_1)}{p(X \mid M_2)}, \tag{2.8}$$

where $p(\mathbf{x} \mid M_k) = \int p(\mathbf{x} \mid \theta_k, M_k)p(\theta_k \mid M_k)\ d\theta_k$. Twice the logarithm of the Bayes factor is approximately equal to the difference between BIC values for the two models being compared (Raftery & Dean, 2006). The BIC approximation of

two models are:

$$BIC_1 = BIC_{clust}(X^C, X^P) \tag{2.9}$$

$$BIC_2 = BIC_{noclust}(X^C) + BIC_{reg}(X^P \mid X^C), \tag{2.10}$$

where $BIC_{clust}(X^C, X^P)$ is the BIC of a GMM with X^P adding useful information to the clustering, $BIC_{noclust}(X^C)$ is the BIC of the current clustering variables, and $BIC_{reg}(X^P \mid X^C)$ is the BIC of the regression of X^P on X^C. Take the difference of BIC_1 and BIC_2, if the difference is greater than zero then add X^P. The proposed variable is added or removed based on the BIC differences of different models.

For the algorithm of the methods, at each step, the method selects the variable that improves the clustering the most. The algorithm firstly selects a variable that has the most evidence of univariate clustering. The next step is to select the second variable which has the most evidence of bivariate clustering. The third step chooses another variable (excluding the previous two) that has the most evidence of multivariate clustering. The fourth step starts to remove the variables from the current selected variables. Comparing multivariate clustering of all selected variables with multivariate clustering of selected variables except the target variable, if the target variable has weak evidence of being included then remove it. The algorithm stops when both the third and fourth steps are rejected.

Chapter 3

Methodology

3.1 Parameter Estimation

3.1.1 Gaussian model-based clustering likelihood

The likelihood is usually used to make inferences on the unknown parameters. Considering cluster analysis, the likelihood for Gaussian model-based clustering has the following formula (McNicholas, 2016)

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{3.1}$$

and the log-likelihood is

$$l(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log \left(\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right). \tag{3.2}$$

It is difficult to estimate the unknown parameters by using this formula. For example, to find the maximum likelihood estimate of μ_g, take the derivative of the

log-likelihood and set the equation equal to 0,

$$\sum_{i=1}^{n} \frac{1}{\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \frac{\mathbf{x_i} - \boldsymbol{\mu_g}}{\sigma_g^2} = 0. \tag{3.3}$$

But it is impossible to solve the above equation. Let Z be the indicator vector such that $Z = (Z_1, ..., Z_n)$. Assume $Z_1, ..., Z_n$ are independent and identically distributed with a multinomial distribution (generalization of the binomial distribution). There are $\mathbf{z}_i = (z_{i1}, ..., z_{iG})$ realized values of the random vectors $Z_1, ..., Z_n$, and they are defined as

$$\mathbf{z}_{ig} = \begin{cases} 1, & \text{if observation } \mathbf{x}_i = (x_{i1}, ..., x_{ip}) \text{ belongs to the } g\text{th group} \\ 0, & \text{other wise} \end{cases} \tag{3.4}$$

where $i \in \mathbb{Z}, i \in [1, n]$.

3.1.2 The EM algorithm

The EM algorithm was proposed by Dempster et al. (1977) and it is widely used for parameter estimation for a finite number of components. The set of pairs $\{(X, Z)\}$ is considered as the complete data set (Bouveyron & Brunet-Saumard, 2014). The following formula is the complete-data likelihood

$$L_c(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left(\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right)^{z_{ig}}, \tag{3.5}$$

and the complete-data log-likelihood is

$$l_c(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left(\log \pi_g + \log \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right). \tag{3.6}$$

In the E-step, the expected value of the complete-data log-likelihood is updated. This is the expected value of z_{ig},

$$\hat{z}_{ig} = \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\mu}_g, \hat{\Sigma}_g)}{\sum_{h=1}^{G} \hat{\pi}_h \phi(\mathbf{x_i} \mid \hat{\mu}_\mathbf{h}, \hat{\mathbf{\Sigma}}_\mathbf{h})}. \tag{3.7}$$

The E-step is conditional on the current parameter estimates. The expected value of the complete-data log-likelihood is

$$E[l_c(\pi,\theta)] = \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}\left(\log \pi_g - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_g| - \frac{1}{2} tr\{(x_i - \boldsymbol{\mu}_g)\boldsymbol{\Sigma}_g^{-1}(x_i - \boldsymbol{\mu}_g)'\}\right). \tag{3.8}$$

In the M-step, the model parameters are updated. By maximizing $E[l_c(\pi,\theta)]$, you get

$$\hat{\pi}_g = \frac{n_g}{n} \tag{3.9}$$

$$\hat{\mu_g} = \frac{1}{n_g}\sum_{i=1}^{n} \hat{z}_{ig}\mathbf{x}_i \tag{3.10}$$

$$\hat{\Sigma}_g = \frac{1}{n_g}\sum_{i=1}^{n} \hat{z}_{ig}(x_i - \hat{\mu}_g)(x_i - \hat{\mu}_g)' \tag{3.11}$$

where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$.

3.1.3 Predicted Clusters

Peel & MacLahlan (2000) suggested the mixing proportion π_g can be considered as the prior probability that observation $\mathbf{x}_i$ belongs to the gth cluster. The posterior probability can be found using the following

$$P(Z_{ig} \mid \mathbf{x}_i) = \frac{\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{h=1}^{G} \pi_h \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}, \tag{3.12}$$

where $h \in [1, ..., G]$, and $E[Z_{ig} \mid \mathbf{x}_i] = P(Z_{ig} = 1 \mid \mathbf{x}_i)$.

After the parameter estimation, each observation is assigned to the corresponding cluster. There is a soft posteriori classification prediction method

$$\hat{z}_{ig} = \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\mu}_g, \hat{\Sigma}_g)}{\sum_{h=1}^{G} \hat{\pi}_h \phi(\mathbf{x}_i \mid \hat{\mu}_h, \hat{\Sigma}_h)} \tag{3.13}$$

(McNicholas, 2016), and there is a hardened posteriori classification method. Sometimes, it is preferable to use a posteriori (MAP) classification method (Peel & MacLahlan, 2000; McNicholas, 2016), where

$$\text{MAP}\{\hat{z}_{ig}\} = \begin{cases} 1, & \text{if } g = \arg\max_h\{\hat{z}_{ih}\} \\ 0, & \text{other wise.} \end{cases} \tag{3.14}$$

3.2 Model Selection

For model selection, the Bayesian information criterion (BIC) is a typical criterion to use (Neath & Cavanaugh, 2012, Schwarz, 1978). Let $M_1, ..., M_m$ be m

candidate models, and each model has a parametric distribution $f_i(x \mid \theta_i)$. The prior distribution is $\pi_i(\theta_i)$, where θ_i contains p parameters, $i = 1, ..., n$. In the book published by Konishi & Kitagawa (2008), the BIC is defined as

$$\begin{aligned} BIC &= -2 \ln L(\hat{\theta} \mid x) + p \ln(n) \\ &= -2l(\hat{\theta} \mid x) + p \ln(n), \end{aligned} \tag{3.15}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ, l is the maximized log-likelihood, p is the total number of parameters, and n is the total number of observations.

3.3 Performance Assessment

There are three critical criteria to determine the efficiency of clustering, "misclustering error" that measures the percentage of off diagonal data (which correspond to mislabeled observations), "performance time" measures the average performance time in seconds, and "adjusted rand index" (ARI).

The Rand index (RI) is the ratio of pairwise agreements to the total number of pairs (Rand, 1971). The Rand index can take values from 0 to 1, when the value is equal to 1 that means it is the perfect clustering (or classification). The drawback of the Rand index is that it is not equal to 0 under random classification. Hubert & Arabie (1985) proposed the adjusted rand index (ARI). Consider partition of the N data samples into two different number of clusters, G clusters and K clusters.

The formula is given as following (Fop & Murphy, 2018)

$$ARI = \frac{\sum_g^G \sum_k^K \binom{N_{gk}}{2} - \left(\sum_g^G \binom{N_{g.}}{2} \sum_k^K \binom{N_{.k}}{2}\right) / \binom{N}{2}}{\frac{1}{2}\left(\sum_g^G \binom{N_{g.}}{2} + \sum_k^K \binom{N_{.k}}{2}\right) - \left(\sum_g^G \binom{N_{g.}}{2} \sum_k^K \binom{N_{.k}}{2}\right) / \binom{N}{2}}, \tag{3.16}$$

where N_{gk} is the number of observations falling in cluster g and k, $N_{g.} = \sum_k N_{gk}$, and $N_{k.} = \sum_g N_{gk}$. The general form is (McNicholas, 2016)

$$ARI = \frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}. \tag{3.17}$$

This study considers the given labels (the assigned labels contained in the datasets) as the true labels. When ARI is 1 that represents perfect clustering (or classification). When it is random classification, the ARI value will be 0. If the ARI value is negative that means the classification is worse than random classification.

Chapter 4

Application

This section compares variable selection for clustering and classification (VSCC) with variable selection for model-based clustering (Clustvarsel). All data were standardized with means equal to 0 and variances equal to 1, and all the categorical variables were removed. Both clustering methods used the mclust package to initialize the algorithms, and models were chosen via the BIC criterion. The following tables are the predicted clustering results of VSCC and clustvarsel on various data sets. ARI values, mis-clustering rate, and performance time(in seconds) were used to assess the efficiency of clustering.

4.1 Banknote Data

The "Banknote" data set (Flury, 1988) from the mclust package contains 100 genuine and 100 counterfeit old-Swiss 1000-franc bank notes. It has 6 measurements, in mm, "length" (length of bill), "left" (width of left edge), "right" (width of right edge), "bottom" (bottom margin width), "top " (top margin width), and "diagonal" (length of diagonal). The correlation plot (Figure 4.1) shows "right"

has a strong positive correlation with "left" (0.74), and "diagonal" has a negative correlation with "bottom" (−0.62). The pairs plot (Figure 4.1) indicates that variable "diagonal" with any other variable provides good discrimination of banknote status. The "bottom" and "top" are important variables for clustering.

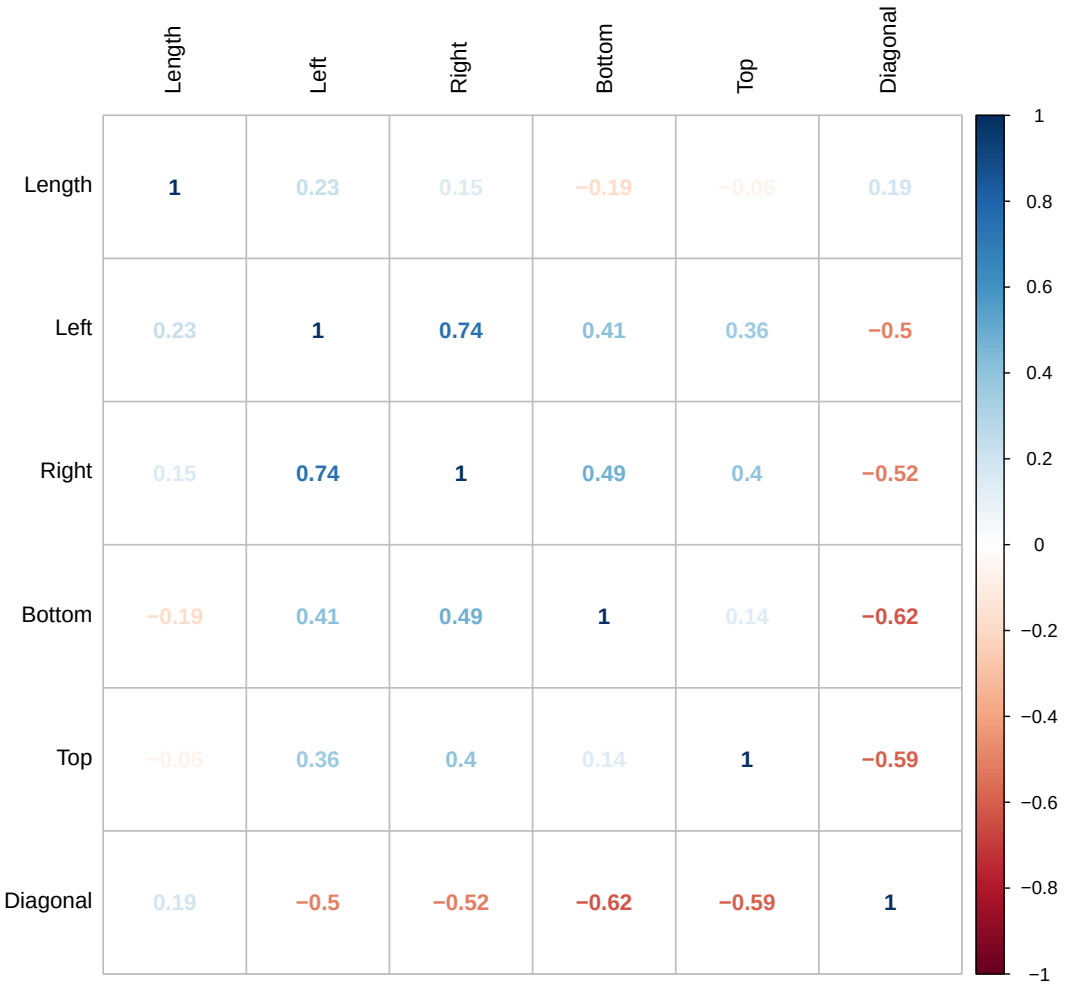

Figure 4.1: The correlation plot of the Banknote data set variables

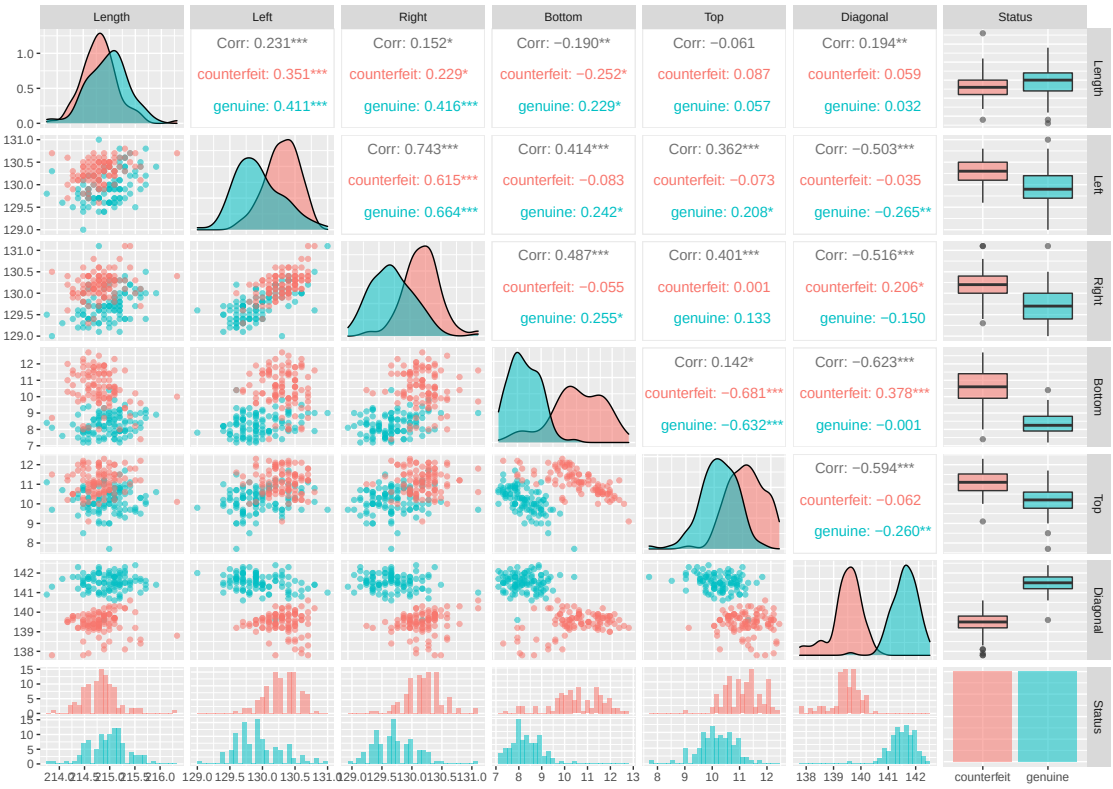

Figure 4.2: The pairs plot of the Banknote data set variables

Table 4.1 and Table 4.2 show the prediction result of banknote data by implementing VSCC and clustvarsel perspectively. Table 4.3 shows that VSCC selects 4 variables out of 6, they are “diagonal”, “bottom”, “top”, and “right”. Clustvarsel selects 4 variables that are “diagonal”, “bottom”, “top”, and “left”. Both of the methods select variable “diagonal”. The VSCC prediction has 3 clusters and clustvarsel has 4 clusters. VSCC has a higher ARI value of 0.8603 as compared to clustvarsel, which has an ARI value of 0.6907. VSCC has an error rate of 43.00% which is smaller than clustvarsel’s at 60.00%. VSCC also has shorter running time in R with 4.93 seconds as compared to clustvarsel with a runtime of 25.43 seconds.

Table 4.1: Prediction table for VSCC

	1	2	3
counterfeit	15	0	85
genuine	1	99	0

Table 4.2: Prediction table for clustvarsel

	1	2	3	4
counterfeit	1	0	15	84
genuine	20	79	1	0

Table 4.3: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Banknote data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.8603	43.00%	3	4	4.93
clustvarsel	0.6907	60.00%	4	4	25.43

4.2 Coffee Data

The "Coffee" data set (Streuli, 1973) includes 43 samples from 29 countries around the world, and each coffee sample belongs to one of the two coffee species, Arabica and Robusta. There are 36 Arabica coffee samples and 7 Robusta coffee samples. There are 12 chemical components, "water", "bean weight", "extract yield", "ph value", "free acid", "mineral content", "fat", "caffeine", "trigonelline", "chlorogenic acid", "neochlorogenic acid", and "isochlorogenic acid". The correlation plot (Figure 4.3) shows the relationship between variables. "Fat" has strong negative correlation with "caffeine" (-0.84), "neochlorogenic acid" (-0.69), and "isochlorogenic acid" (-0.67); but has strong positive correlation with "trigoneline" (0.70). "Caffeine" has a strong positive correlation with "isochlorogenic acid" (0.80). Based on the pairs plot (Figure 4.4), "fat" and "caffeine" are significant clustering variables that could identify coffee into two different clusters.

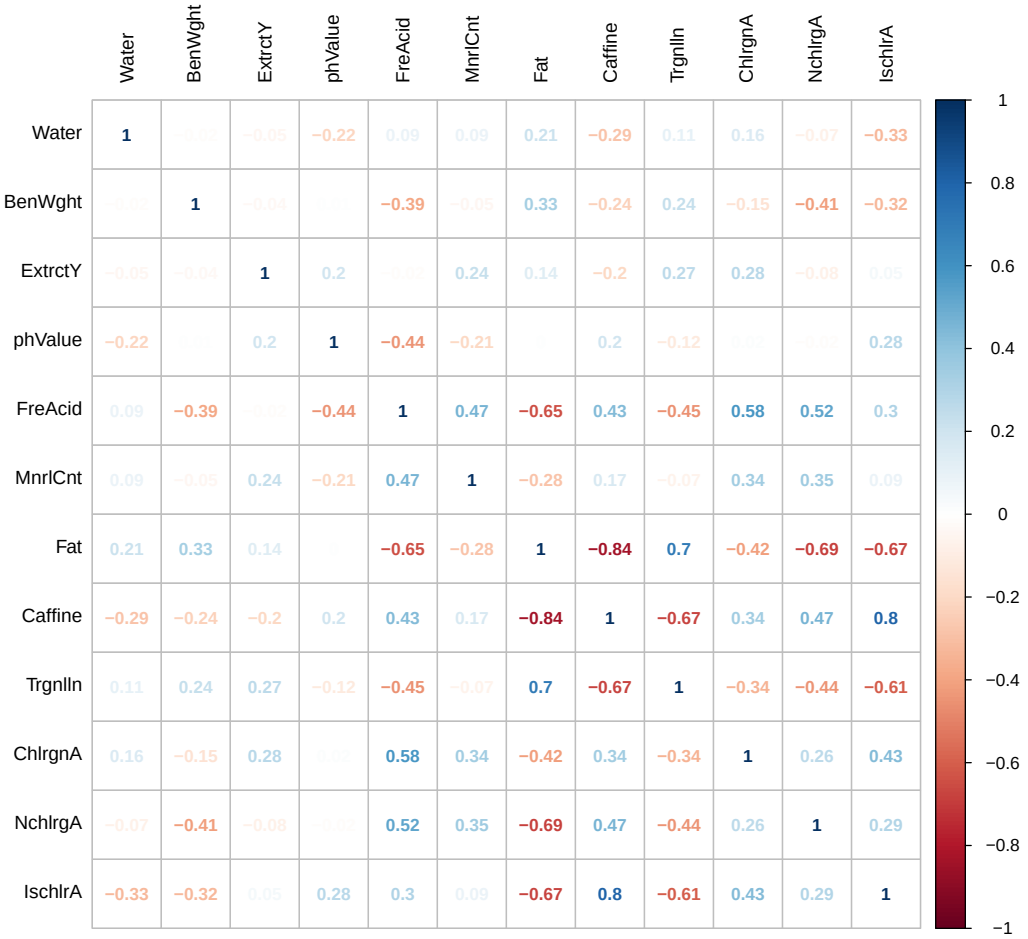

Figure 4.3: The correlation plot of the Coffee data set variables

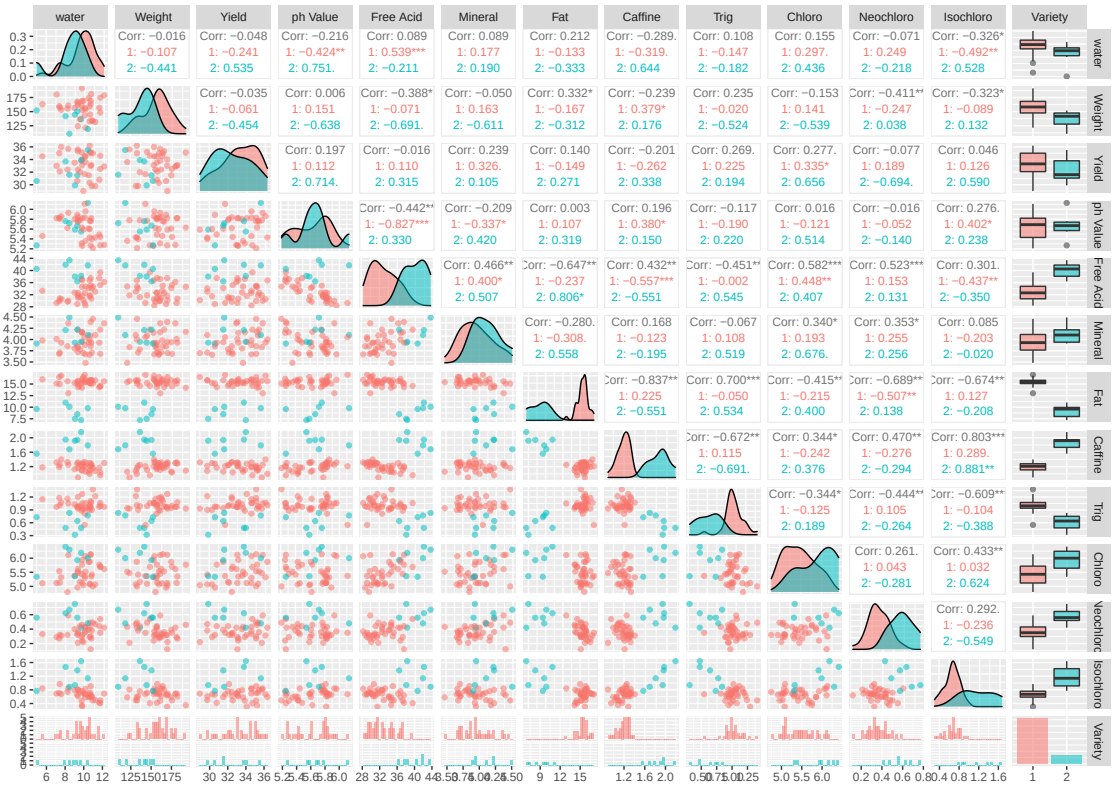

Figure 4.4: The pairs plot of the Coffee data set variables

The following tables (Table 4.4, Table 4.5) are the prediction tables for the coffee, Table 4.4 shows the result of predictions using VSCC, and Table 4.5 shows the result of predictions using clustvarsel. Table 4.6 indicates that VSCC selects 2 variables out of 12, and they are "fat" and "free acid". Clustvarsel selects 5 out of 12 variables and they are "fat", "caffeine", "free acid", "ph value", and "extract yield". G represents the number of clusters, VSCC has 2 predicted clusters, and clustvarsel has 3. VSCC has an ARI value of 1, which is higher than clustvarsel's ARI value of 0.3732, and VSCC also has a smaller clustering error than clustvarsel. The performance time for VSCC is shorter than for clustvarsel.

Table 4.4: Prediction table for VSCC

	1	2
1	36	0
2	0	7

Table 4.5: Prediction table for clustvarsel

	1	2	3
1	21	15	0
2	0	0	7

Table 4.6: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Coffee data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	1	0%	2	2	1.56
clustvarsel	0.3732	51.16%	3	5	10.95

4.3 Glass Identification Data

The "Glass Identification" data set (Evett & Ernest, 1987) has 214 samples and 9 variables. There are 7 types of glass, class 1 represents building windows float processed, class 2 represents building windows non-float processed, class 3 and class 4 represent vehicle windows float processed and vehicle windows non-float processed respectively (but there is no type 4 glass samples in this data set), class 5 represents containers, class 6 represents tableware, and class 7 represents headlamps. The variable "RI" is refractive index. Besides glass id number and glass type, the other variables are 8 chemical components, "Na" (sodium, weight percent in corresponding oxide), "Mg" (magnesium), "Al" (aluminium), "Si" (silicon), "K" (Potassium), "Ca" (Calcium), "Ba" (barium), and "Fe" (iron). The correlation plot, Figure 4.5 shows "Si" is negatively correlated with "RI" (-0.54), and "Ba" is negatively correlated with "Mg" (-0.49). "Ca" has a strong positive

correlation with “RI” (0.81), and a weak negative correlation with “Mg” (−0.44).

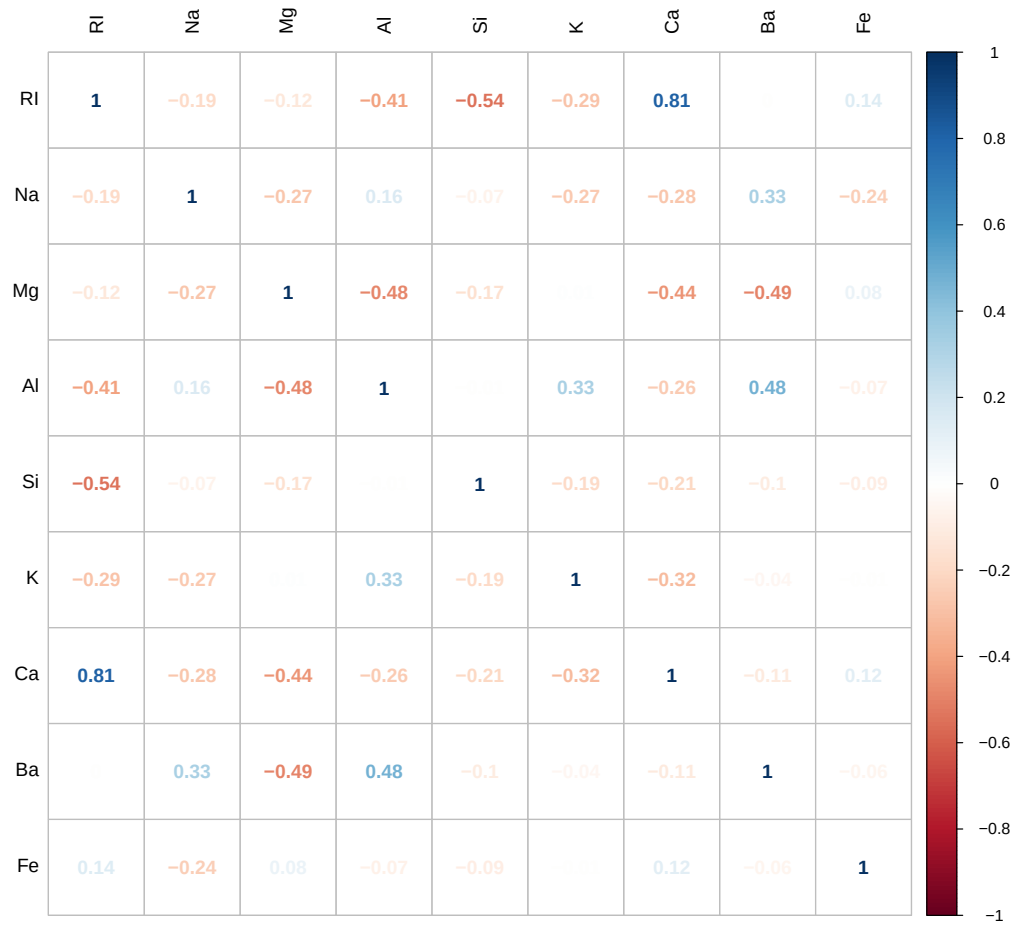

Figure 4.5: The correlation plot of the Glass data set variables

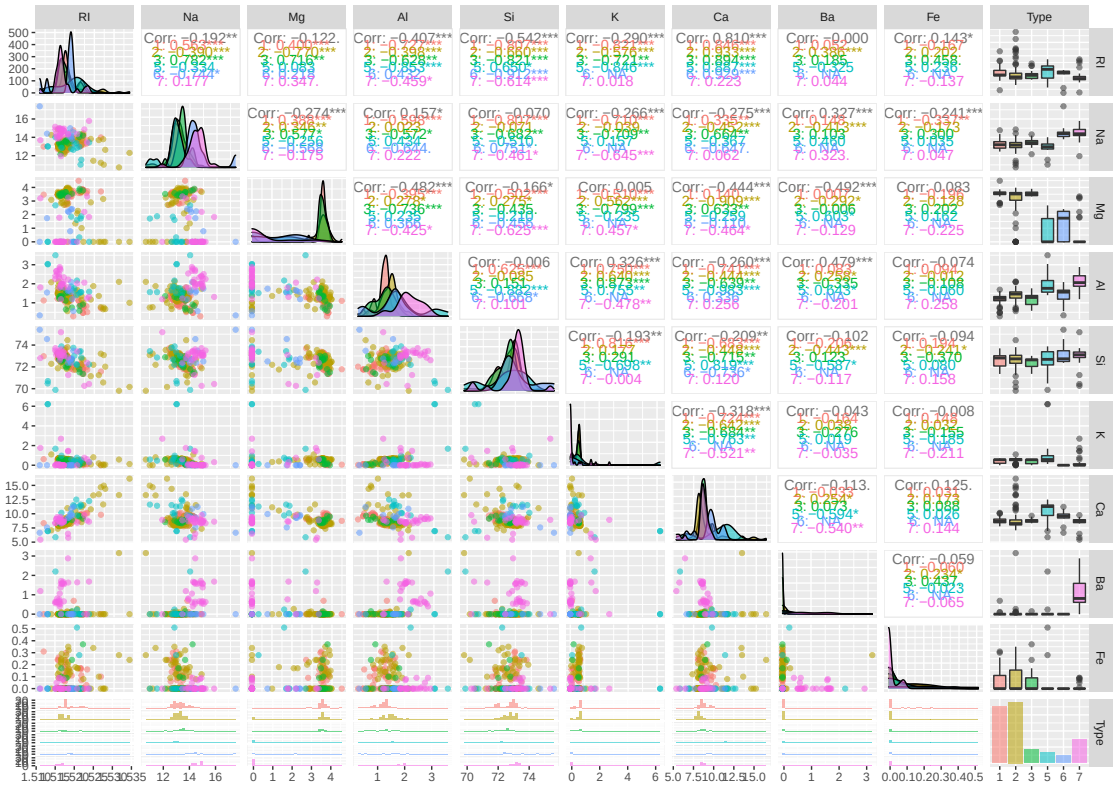

Figure 4.6: The pairs plot of the Glass data set variables

The VSCC prediction table (Table 4.7) shows there are 5 predicted types of glass, and the clustvarsel prediction table (Table 4.8) has two predicted classes. Table 4.9 indicates that VSCC has a higher ARI value of 0.1470 than the clustvarsel ARI of 0.1466, but VSCC has a 3.27% higher clustering error rate than clustvarsel. The VSCC method selects 9 variables, "RI", "Na", "Mg", "Al", "Si", "K", "Ca", "Ba", and "Fe"; and clustvarsel selects "Ba" and "Al". The running time for VSCC is approximately 2 seconds shorter than clustvarsel.

Table 4.7: Prediction table for VSCC

	1	2	3	5	7
1	14	32	21	3	0
2	11	35	18	12	0
3	3	8	6	0	0
5	2	0	0	11	0
6	6	0	0	3	0
7	0	1	0	8	20

Table 4.8: Prediction table for clustvarsel

	1	2
1	67	3
2	70	6
3	16	1
5	11	2
6	9	0
7	3	26

Table 4.9: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Glass data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.1470	69.16%	5	9	3.14
clustvarsel	0.1465	65.89%	2	2	5.05

4.4 Wine Recognition Data

The "Wine Recognition" data set (Blake & Merz, 1998) from the gclus package contains 178 samples of Italian wines from three different cultivars. A chemical analysis of these wines yielded 13 measurements, "alcohol", "malic" (malic acid), "ash", "alcalinity" (alcalinity of ash), "magnesium", "phenols" (total phenols), "flavanoids", "nonflavanoid", "proanthocyanins", "intensity", "Hue", "OD280", and "proline". The correlation plot (Figure 4.7) shows "flavanoids" has a strong positive correlation with "phenols" (0.86) and OD280 (0.79). The pairs plot (Figure 4.8) indicates that proline and intensity are significant clustering variables for

the wine data as well.

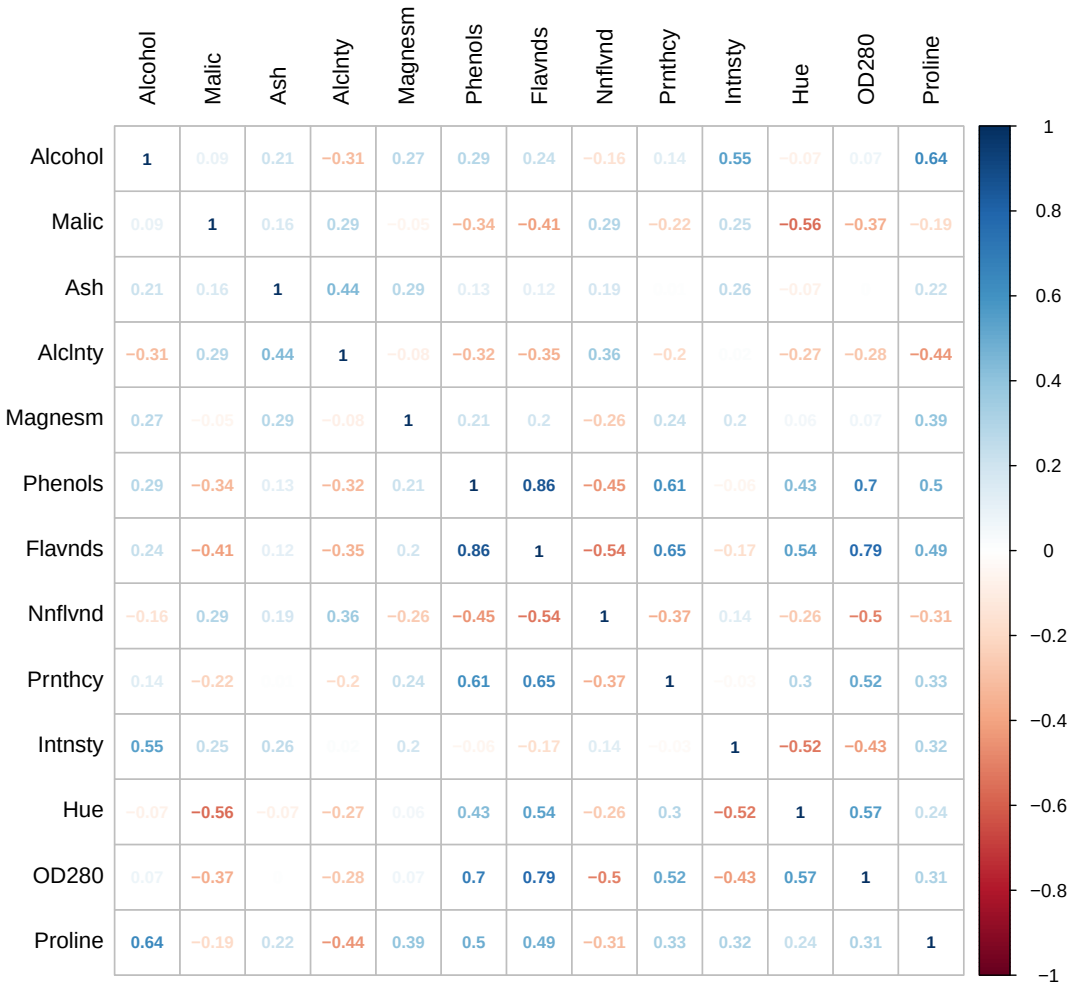

Figure 4.7: The correlation plot of the Wine Recognition data set variables

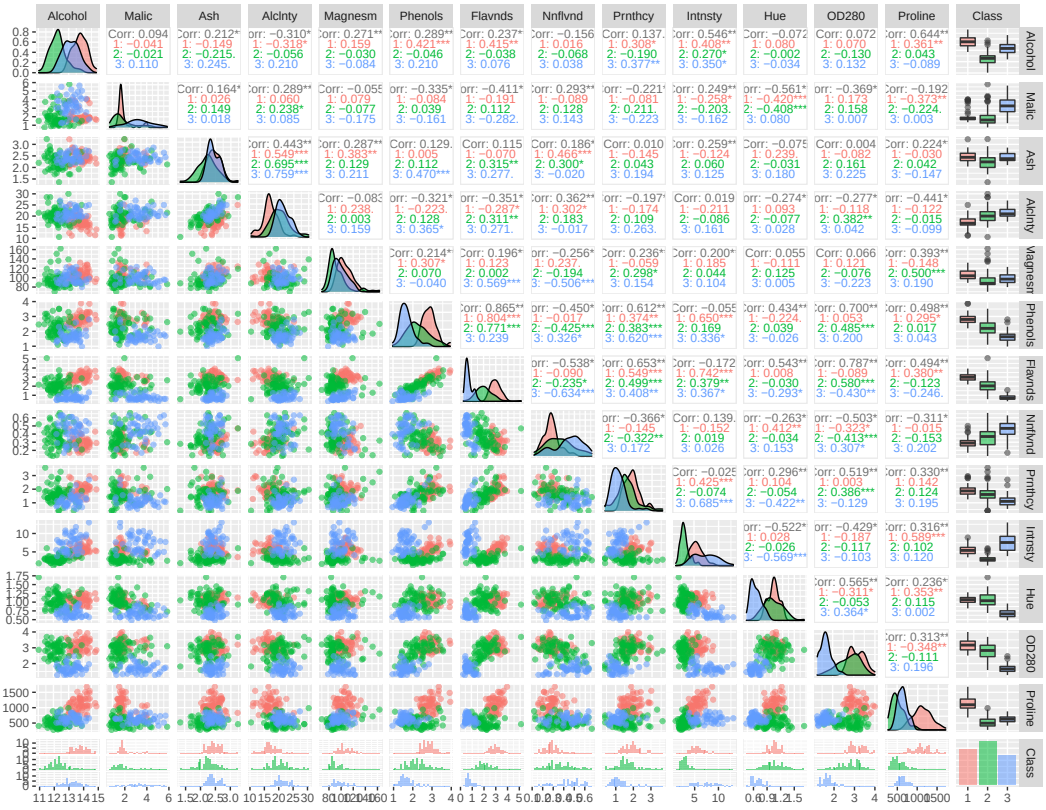

Figure 4.8: The pairs plot of the Wine Recognition data set variables

The following two tables (Table 4.10 and Table 4.11) are the predicted cluster memberships with VSCC and clustvarsel methods, with both of the methods having 3 predicted clusters. VSCC has a higher ARI value (0.9297) than clustvarsel (0.7828). The clustering error for VSCC is 2.25% and for clustvarsel is 7.30%. VSCC selects 13 variables, and clustvarsel selects 5 variables that are malic, proline, flavanoids, intensity, and OD280. Also, VSCC performs much faster than clustvarsel on the wine recognition data.

Table 4.10: Prediction table for VSCC

	1	2	3
1	56	3	0
2	0	70	1
3	0	0	48

Table 4.11: Prediction table for clustvarsel

	1	2	3
1	51	8	0
2	3	67	1
3	0	1	47

Table 4.12: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Wine Recognition data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.9297	2.25%	3	13	6.87
clustvarsel	0.7828	7.30%	3	5	62.83

4.5 Iris Data

The "Iris" data set (Becker et al., 1988) has 150 samples in total with 50 flowers from each of 3 species of iris, the species are "Iris setosa", "versicolor", and "virginica". Figure 4.10 shows there are 4 measurements in centimetres of the variables "sepal length", "sepal width", "petal length" and "petal width". Figure 4.9 shows that "sepal length" has strong positive correlation with "petal length" (0.87) and "petal width" (0.82). "Petal length" and "width" have a high positive correlation at 0.96.

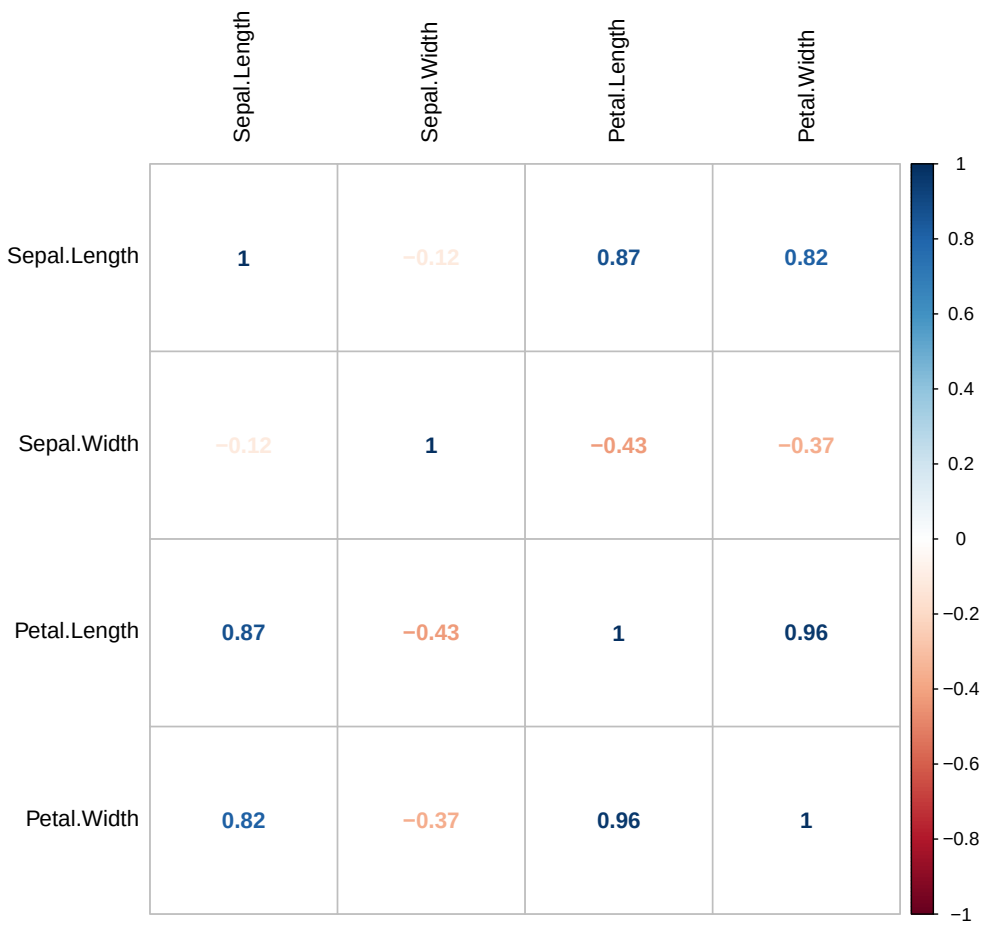

Figure 4.9: The correlation plot of the Iris data set variables

Figure 4.10: The pairs plot of the Iris data set variables

Table 4.13 and Table 4.14 are the prediction tables for the Iris data, VSCC predicts 2 clusters and clustvarsel predicts 3 clusters. VSCC selects all the variables and clustvarsel selects petal length, petal width, and sepal width. Table 4.15 shows clustvarsel has a higher ARI (0.7196) than VSCC (0.5681), and it also has a lower mis-clustering error at 11.44% than VSCC at 44.44%. VSCC has a shorter performance time than clustervarsel.

Table 4.13: Prediction table for VSCC

	1	2
setosa	50	0
versicolor	0	50
virginica	0	50

Table 4.14: Prediction table for clustvarsel

	1	2	3
setosa	50	0	0
versicolor	0	50	0
virginica	0	17	33

Table 4.15: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Iris data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.5681	33.33%	2	4	2.91
clustvarsel	0.7196	11.33%	3	3	8.79

4.6 Italian Olive Oil Data

The "Italian olive oil" data set (Forina & Tiscornia, 1982; Forina et al., 1983; Swayne et al., 2006) contains information on the percentage composition of 8 fatty acids found by lipid fraction of 572 samples. These samples are from 3 regions: Southern Italy, Sardinia, and Northern Italy. There are 323 olive oil samples from Southern Italy, 98 samples from Sardinia, and 151 from Northern Italy. Each region has a different number of areas. "Southern Italy" has 4 areas "North Apulia", "Calabria", "South Apulia", and "Sicily"; "Sardinia" comprises 2 areas "Inland Sardinia" and "Costal Sardinia"; and "Northern Italy" comprises 3 areas "East Liguria", "West Liguria", and "Umbria". The data set uses numbers 1 to 9 to represent these regions respectively. Figure 4.12 shows the 8 fatty acids, "Palmitic acid", "Palmitoleic acid", "Stearic acid", "Oleic acid", "Linoleic acid",

"Linolenic acid", "Arachidic acid", and "Eicosenoic acid". Figure 4.11 indicates that "Palmitic acid" has a strong positive correlation with "Palmitoleic acid" at 0.84 and a strong negative correlation with "Oleic acid" at −0.84. "Oleic acid" has strong negative correlation with "Palmitoleic acid" at −0.85 and "Linoleic acid" at −0.85.

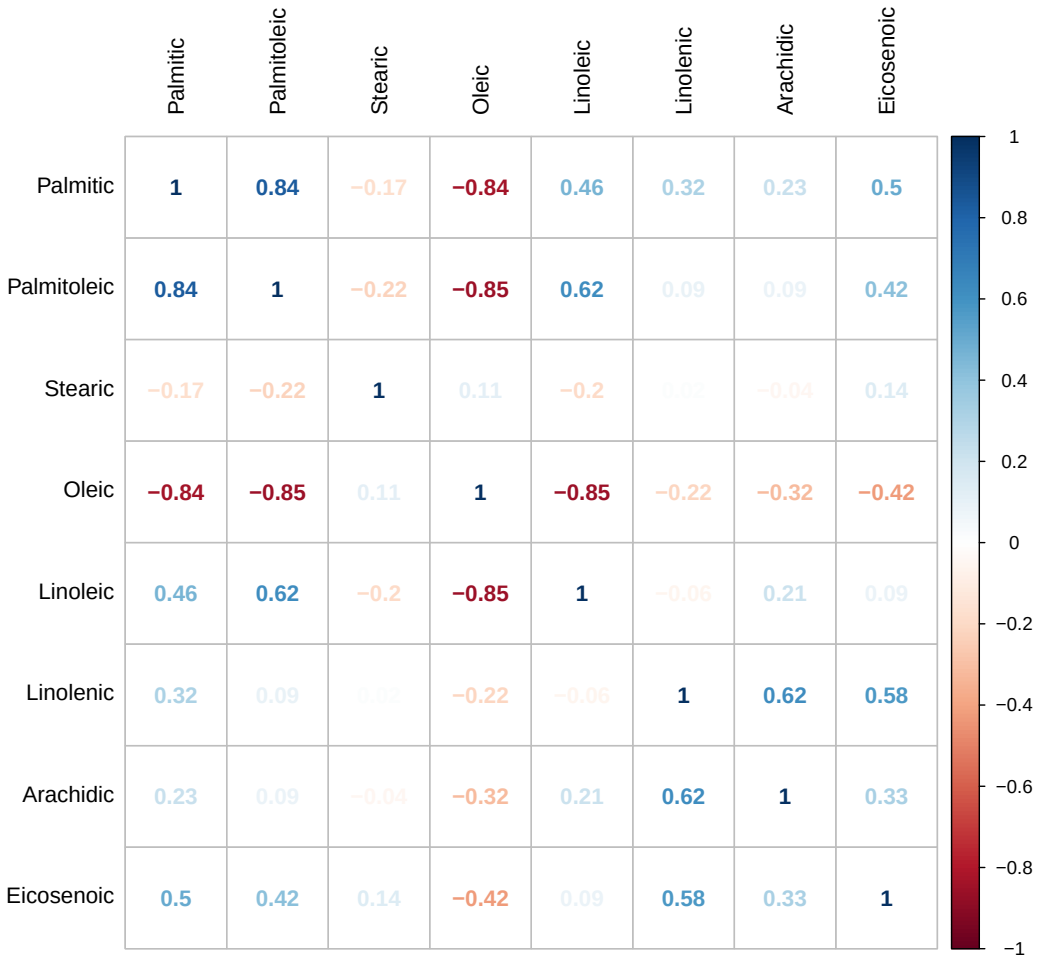

Figure 4.11: The correlation plot of the Italian Olive Oil data set variables

Figure 4.12: The pairs plot of the Italian Olive Oil data set variables

The following two tables (Table 4.16 and Table 4.17) show the prediction labels of the "Italian Olive Oil" data set for VSCC and clustvarsel. Two methods have the same prediction tables. Table 4.18 shows both of the methods select 8 variables (all the variables) and have 9 predicted clusters. The two methods have the same ARI at 0.6586 and the same mis-clustering error at 31.29% but VSCC has a shorter performance time at 11.49 seconds.

Table 4.16: Prediction table for VSCC

	1	2	3	4	5	6	7	8	9
North Apulia	23	2	0	0	0	0	0	0	0
Calabria	0	56	0	0	0	0	0	0	0
South Apulia	0	6	79	121	0	0	0	0	0
Sicily	6	29	1	0	0	0	0	0	0
Inland Sardinia	0	0	0	0	65	0	0	0	0
Coastal Sardinia	0	0	0	0	2	31	0	0	0
East Liguria	0	0	0	0	0	0	49	1	0
West Liguria	0	0	0	0	0	0	3	47	0
Umbria	0	0	0	0	0	0	8	0	43

Table 4.17: Prediction table for clustvarsel

	1	2	3	4	5	6	7	8	9
North Apulia	23	2	0	0	0	0	0	0	0
Calabria	0	56	0	0	0	0	0	0	0
South Apulia	0	6	79	121	0	0	0	0	0
Sicily	6	29	1	0	0	0	0	0	0
Inland Sardinia	0	0	0	0	65	0	0	0	0
Coastal Sardinia	0	0	0	0	2	31	0	0	0
East Liguria	0	0	0	0	0	0	49	1	0
West Liguria	0	0	0	0	0	0	3	47	0
Umbria	0	0	0	0	0	0	8	0	43

Table 4.18: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Italian Olive Oil data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.6586	31.29%	9	8	11.49
clustvarsel	0.6586	31.29%	9	8	177.86

4.7 Leptograpsus Crabs Data

The "Leptograpsus Crabs" data set (Ripley, 2002) from the MASS library in R has 200 samples and consists of 50 crabs from each species (blue and orange) and both gender (female and male). The pairs plot (Figure 4.14) shows there are 5 morphologic measurements on two species of crabs in mm, "frontal lobe size" (FL), "rear width" (RW), "carapace length" (CL), "carapace width" (CW), and "body depth" (BD). The correlation plot (Figure 4.13) shows that all the variables have a high positive correlation with each other and are greater or equal to 0.89.

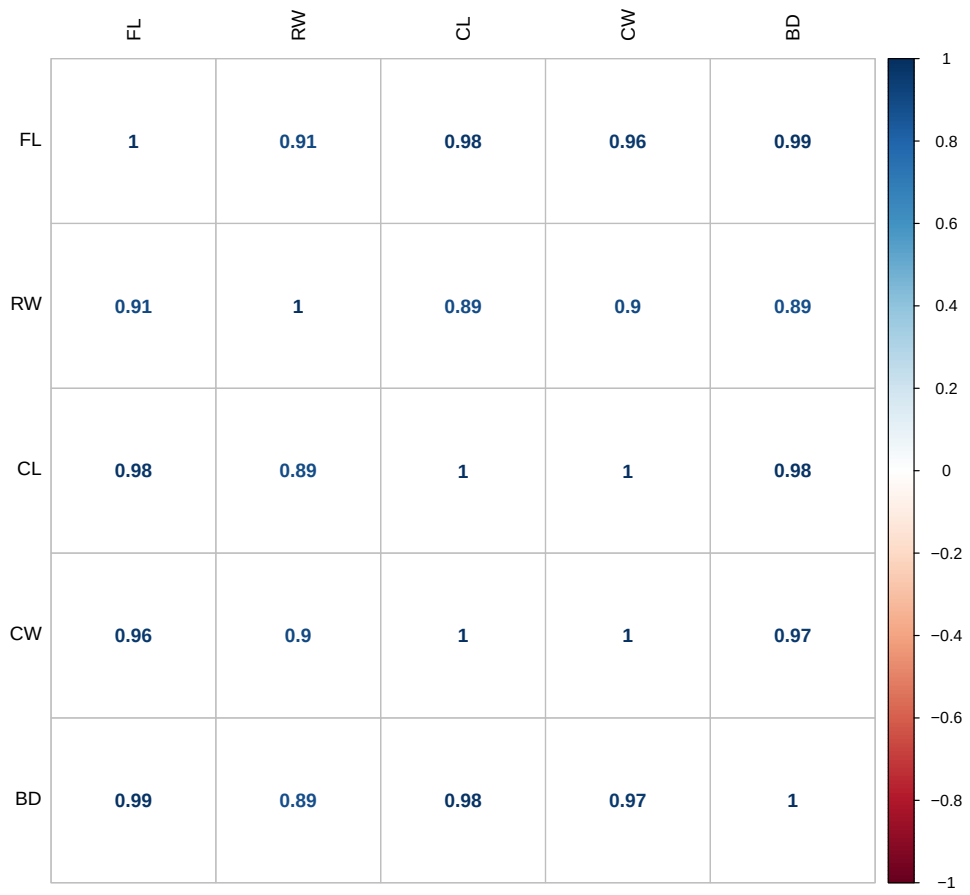

Figure 4.13: The correlation plot of the Leptograpsus Crabs data set variables

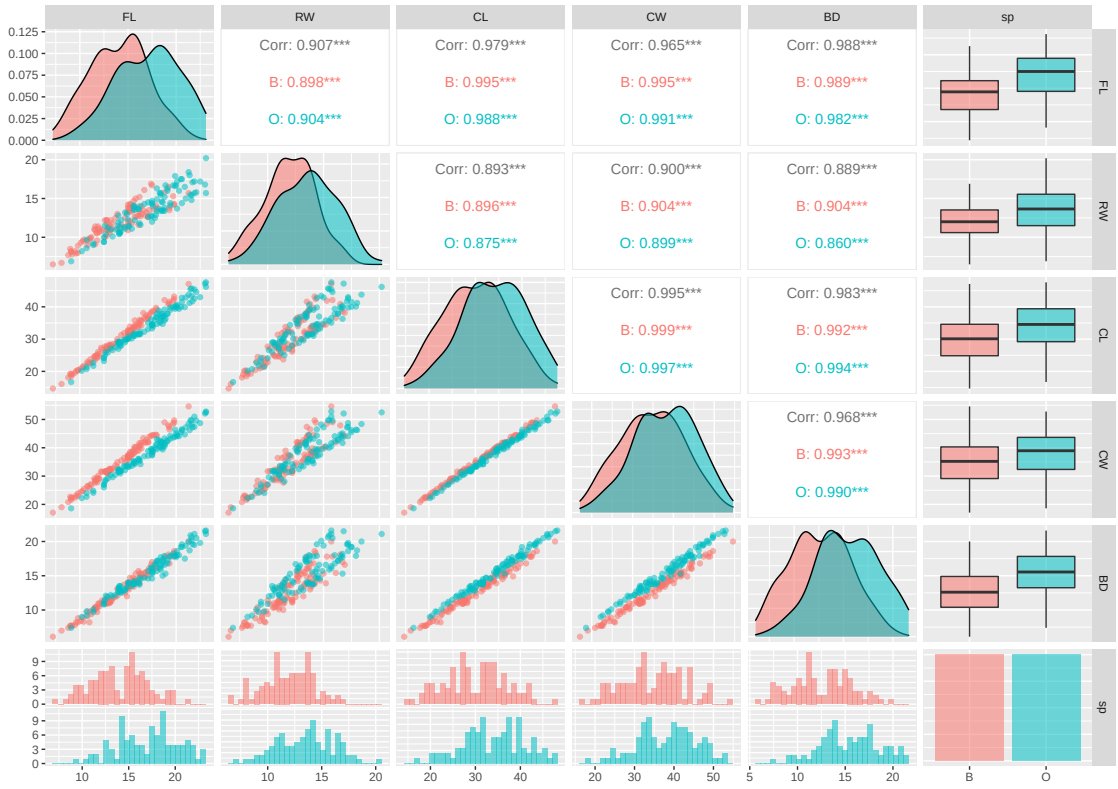

Figure 4.14: The pairs plot of the Leptograpsus Crabs data set variables

The following tables (Table 4.19 and Table 4.20) show the prediction of crab species using VSCC and clustvarsel, where "B" and "O" represent the colours blue and orange, and "M" and "F" represent sexes male and female. Both of the methods cluster the crabs by colour and gender, giving 4 clusters in total. Table 4.21 shows VSCC selects 3 variables that are "frontal lobe size" (FL), "rear width" (RW), and "carapace width" (CW). The method clustvarsel selects 4 variables, 3 are the same as VSCC, in addition to "body depth" (BD). The method clustvarsel has a higher ARI at 0.8291 than VSCC at 0.8052. The mis-clustering error of clustvarsel is 7% which is 1% lower than for VSCC. VSCC has a shorter execution time than clustvarsel.

Table 4.19: Prediction table for VSCC

	BM	OM	BF	OF
BF	50	0	0	0
BM	12	38	0	0
OF	1	0	47	2
OM	0	0	1	49

Table 4.20: Prediction table for clustvarsel

	OM	BM	BF	OF
BF	49	1	0	0
BM	10	40	0	0
OF	0	0	47	3
OM	0	0	0	50

Table 4.21: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Leptograpsus Crabs data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.8052	8%	4	3	5.73
clustvarsel	0.8291	7%	4	4	25.97

4.8 Wheat Kernels Data

The "Wheat Kernels" data set (Charytanowicz et al., 2010) contains three different varieties of wheat: Kama, Rosa, and Canadian. Each type has 70 observations. The experiment used a soft X-ray technique to visualize the internal kernel structure without destruction. There are 7 geometrical properties of kernels that were measured: "area" (A), "perimeter" (P), "compactness" (C), "length of kernel" (LK), "width of kernel" (WK), "asymmetry coefficient" (A_Coef), and "length of kernel groove" (LKG). The correlation plot (Figure 4.15) shows that "area" has strong positive correlation with "perimeter" (0.99), "length of kernel" (0.95), "width of kernel" (0.97), and "length of kernel groove" (0.86). "Perimeter" has strong positive correlation with "length of kernel" (0.97), "width of kernel"

(0.94), and "length of kernel groove" (0.89). "Length of kernel" has positive correlation with "width" at 0.86 and "length of kernel groove" at 0.93. "Width of kernel" and "length of kernel groove" have positive correlation at 0.75. The pairs plot (Figure 4.16) shows the distribution and correlation of seven variables. "Area", "perimeter" and "length of kernel groove" are important variables for clustering.

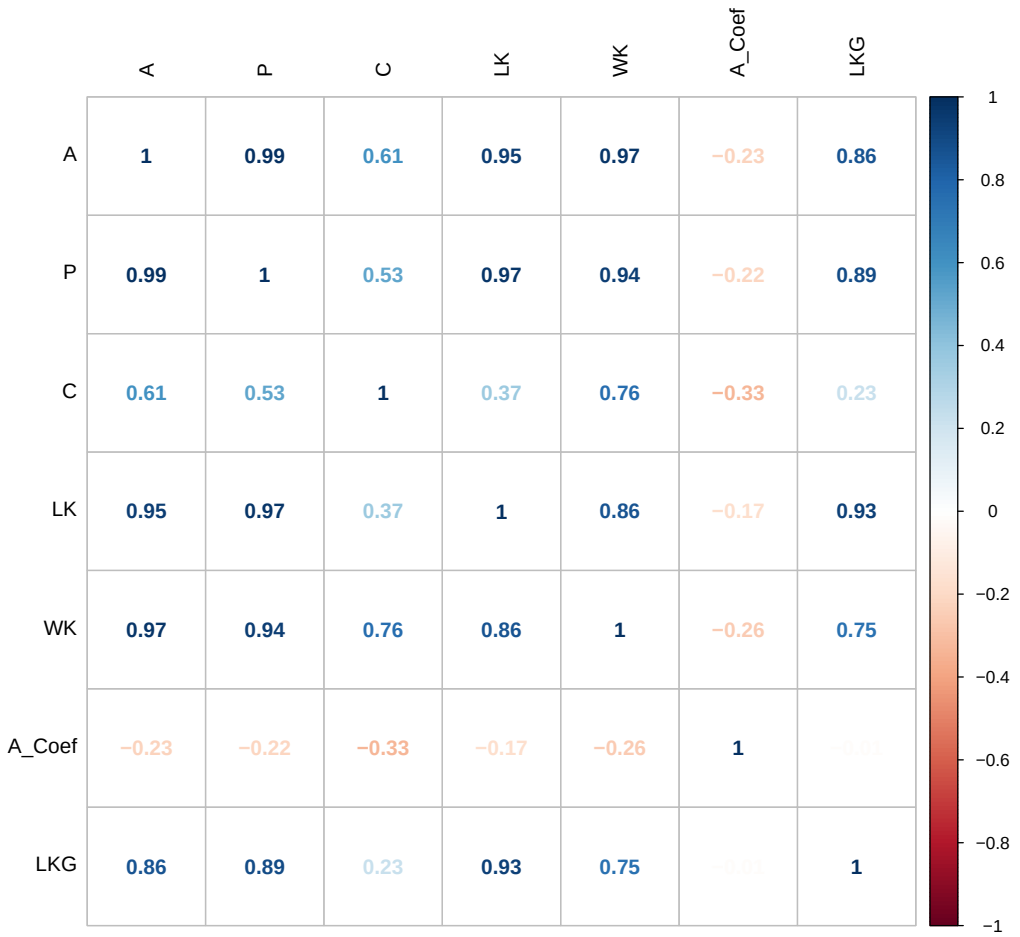

Figure 4.15: The correlation plot of the Wheat Kernel data set variables

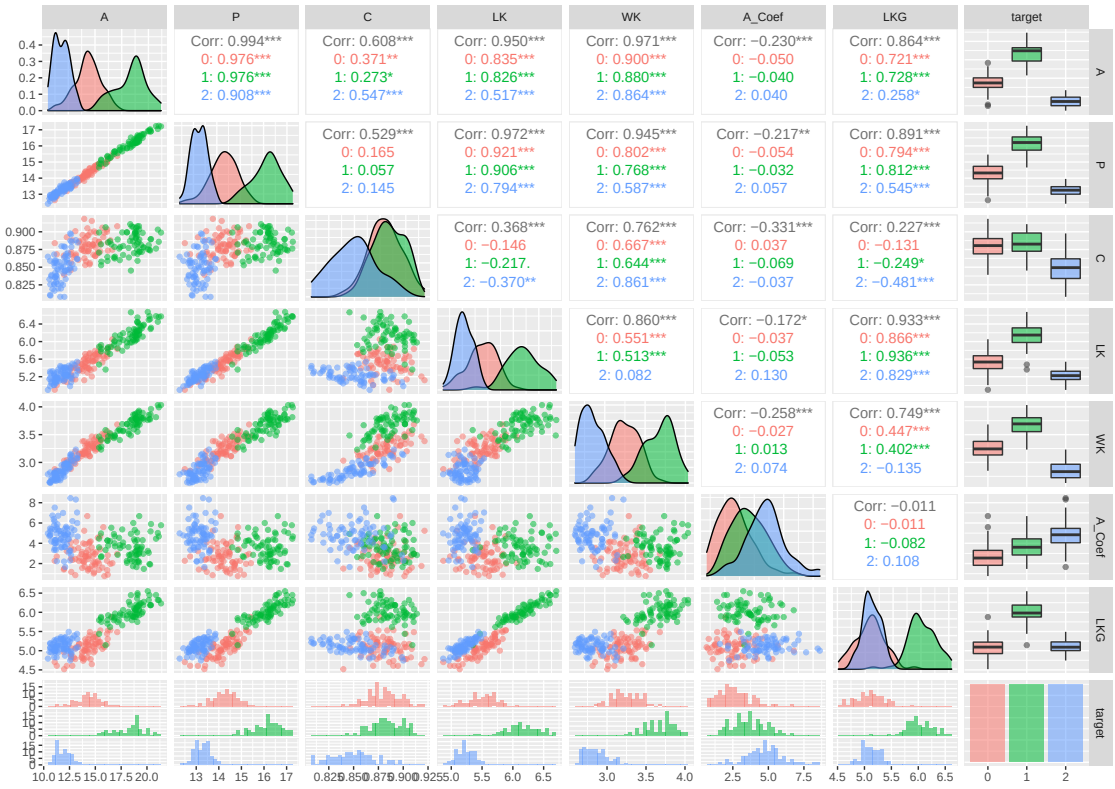

Figure 4.16: The pairs plot of the Wheat Kernel data set variables

Table 4.22 and Table 4.23 are the prediction tables for the wheat kernels by applying VSCC and clustvarsel methods respectively. The VSCC method has two prediction clusters and 5 selected variables, that are “area”, “perimeter”, “width of kernel”, “length of kernel”, and “length of kernel groove”. The clustvarsel has 3 clusters in the prediction table and 3 selected variables, that are “length of kernel groove”, “perimeter”, and “length of kernel”. The clustvarsel method has a higher ARI value of 0.8520 than the VSCC ARI value, 0.5075. Clustvarsel also has a lower prediction error at 5.24% than VSCC at 38.10%. VSCC has a faster running time than clustvarsel.

Table 4.22: Prediction table for VSCC

	1	2
1	60	10
2	0	70
3	70	0

Table 4.23: Prediction table for clustvarsel

	1	2	3
1	64	2	4
2	0	70	0
3	5	0	65

Table 4.24: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Wheat Kernels data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.5075	38.10%	2	5	7.18
clustvarsel	0.8520	5.24%	3	3	26.01

4.9 Wisconsin Breast Cancer Data

The "Wisconsin Breast Cancer" data set is located in the UCI Machine Learning Repository and consists of 569 samples containing 32 variables. The "diagnosis" variable is a binary variable that describes status malignant, or benign. Ten real-valued features are computed for each cell nucleus: "radius" (mean of distances from center to points on the perimeter), "texture" (standard deviation of gray-scale values), "perimeter", "area", "smoothness" (local variation in radius lengths), "compactness" (perimeter^2/ area -1.0), "concavity" (severity of concave portions of the contour), "concave points" (number of concave portions of the contour), "symmetry", and "fractal dimension". For each image, mean, standard error, and worst or largest value are calculated for each image. These features are computed from a digitized image of a fine needle aspiration (FNA) of a breast

mass, and they describe characteristics of the cell nuclei present in the image. The correlation plot (Figure 4.17) shows "mean area" has strong positive correlation with "mean radius" and "mean perimeter". "Worst radius" has strong positive relation with "mean radius", "mean area", "mean perimeter", "worst perimeter", and "worst area". Figures 4.18, 4.19, and 4.20 are the pairs plots of 30 variables. Concavity mean, Concavity points mean, radius se, radius worst, perimeter worst, and concavity points worst are significant variables for clustering. In the parallel coordinate plots (Figure 4.21 and Figure 4.22), number 1 to 30 represent the mean, standard error, and the worst value of the ten measurements of cell nucleus. Figure 4.21 does not have scales and Figure 4.22 standardizes the vertical height and centers each variable at their median. The two parallel coordinate plots indicate variables "radius mean", "perimeter mean", "area mean", "smoothness mean", "symmetry mean", "radius se", "perimeter se", "area se", "concavity se", "fractal dimension se", "radius worst", "perimeter worst", "area worst", "compactness worst", and "fractal dimension worst" are significant variables for clustering.

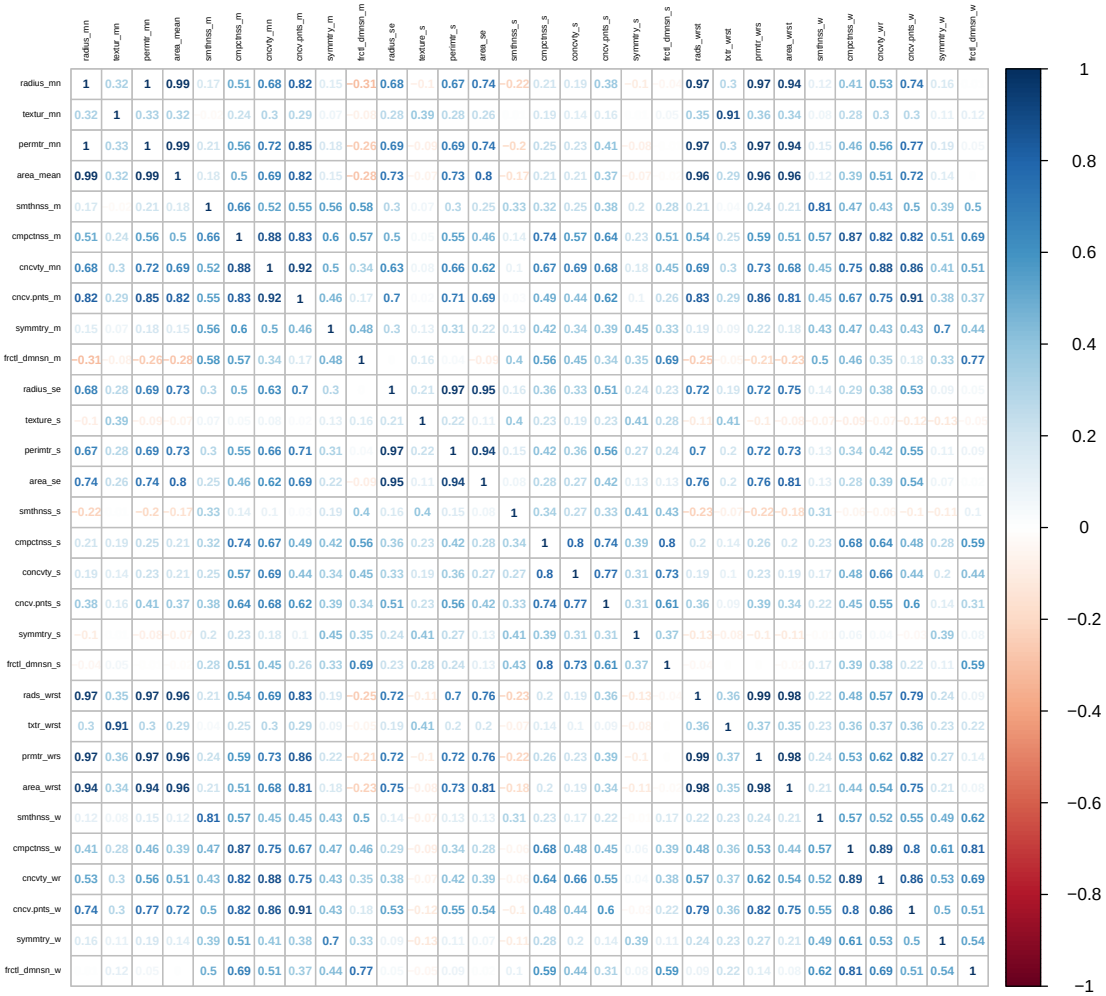

Figure 4.17: The correlation plot of the Wisconsin Breast Cancer data set variables

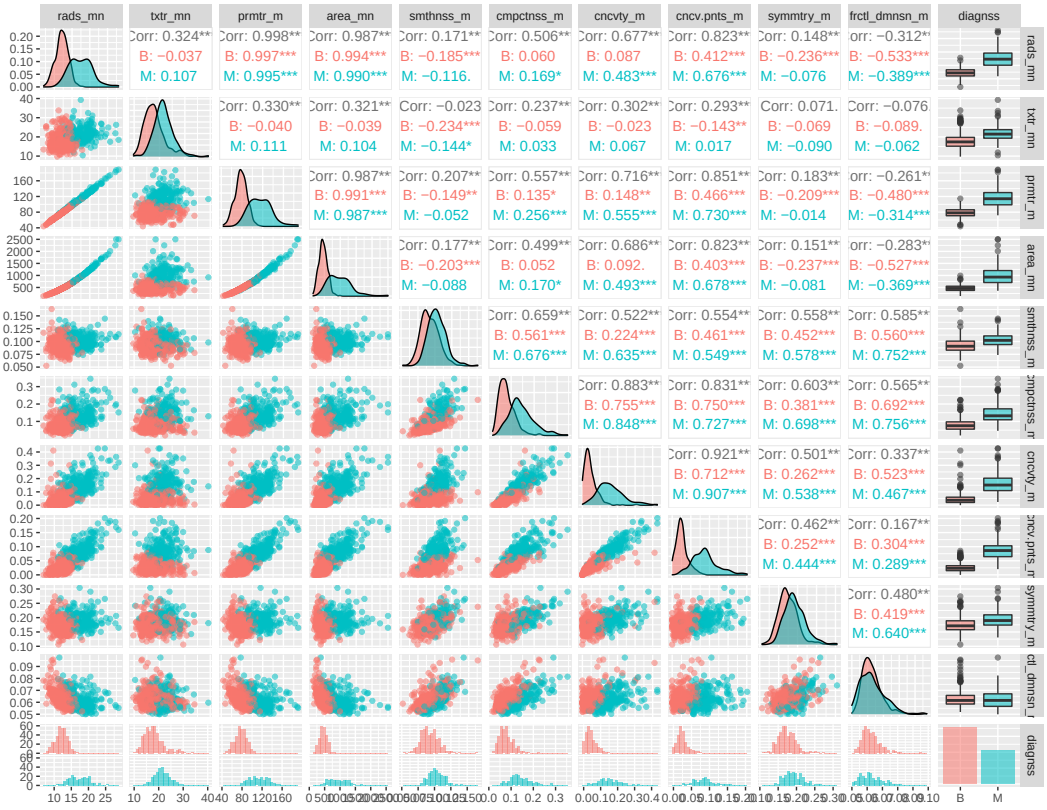

Figure 4.18: The pairs plot of the Wisconsin Breast Cancer data set (mean) variables

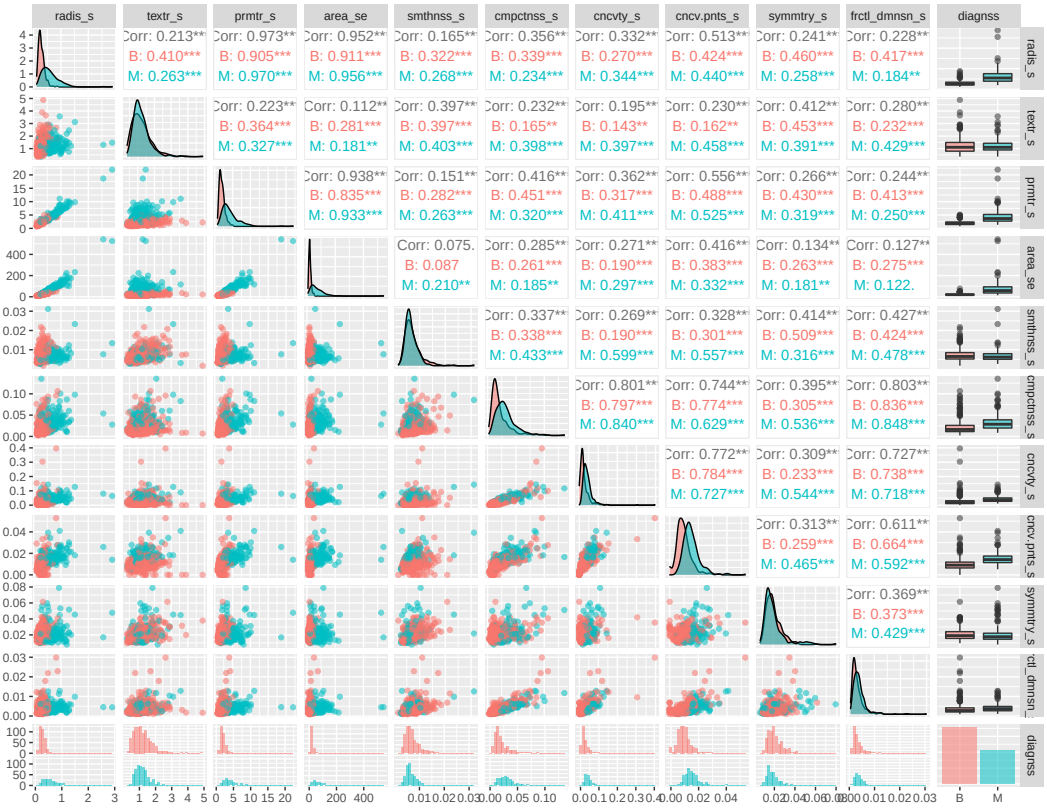

Figure 4.19: The pairs plot of the Wisconsin Breast Cancer data set (standard error) variables

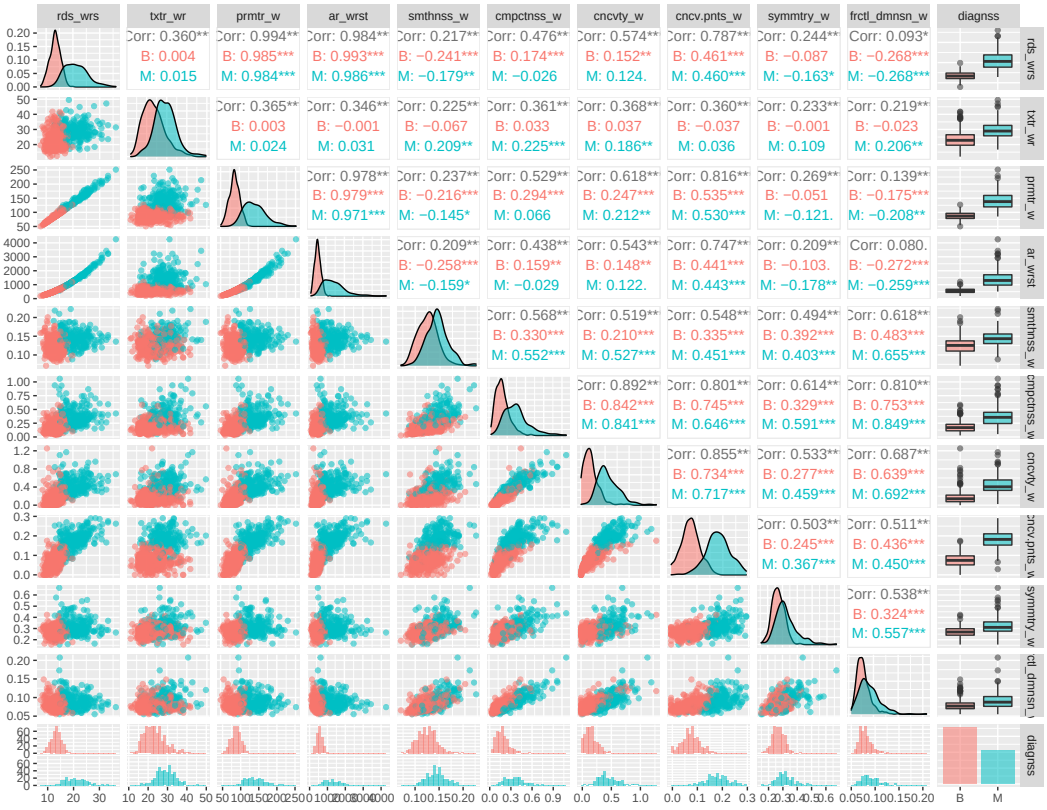

Figure 4.20: The pairs plot of the Wisconsin Breast Cancer data set (worst value) variables

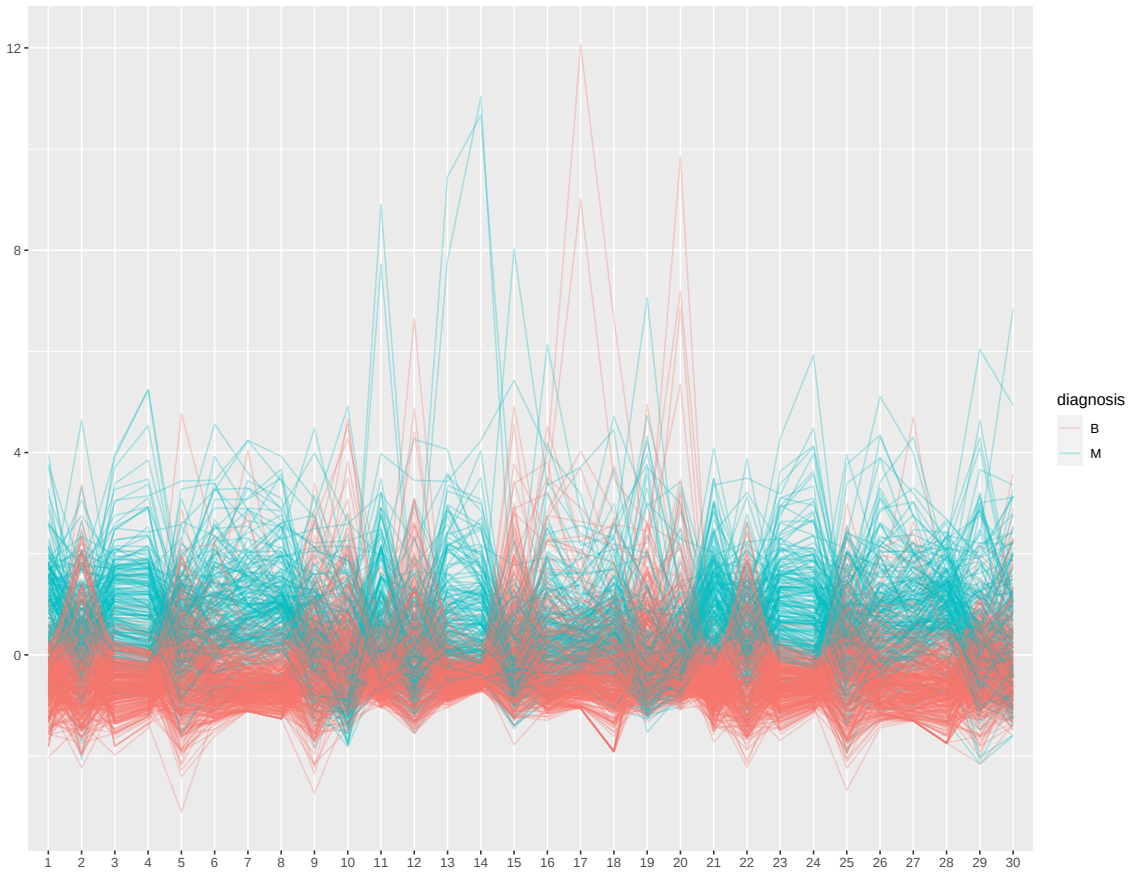

Figure 4.21: The parallel coordinate plot (no scaling) of the Wisconsin Breast Cancer data set variables

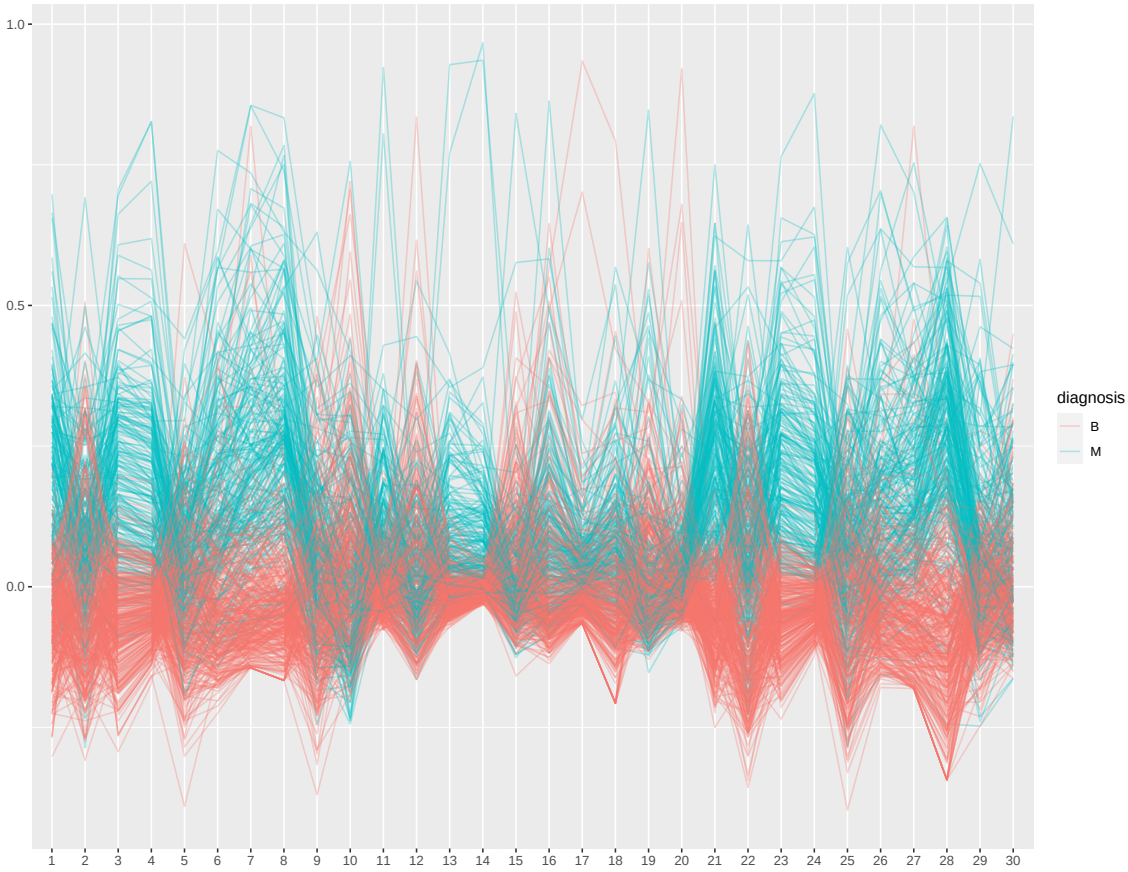

Figure 4.22: The parallel coordinate plot (standardize and center variables) of the Wisconsin Breast Cancer data set variables

Table 4.25 and Table 4.26 are the predication tables for breast cancer data by implementing VSCC and clustvarsel respectively. Table 4.27 indicates that VSCC selects 12 variables out of 30, which are "concavity mean", "compactness mean", "concave points mean", "compactness se", "radius se", "concave points se", "fractal dimension se", "concavity se", "compactness worst", "concave points worst", "concavity worst", and "area worst". The clusvarsel method selects 6 variable that are "area se", "radius se", "area worst", "radius worst", "area mean", and "radius mean". Using a pre-determined cluster size of 2, Table 4.27 shows the clustvarsel ARI value is 0.7060 which is higher than the VSCC ARI value at 0.5254,

but VSCC runs faster than clustvarsel. Also, the mis-clustering error for VSCC is approximately 5% higher.

Table 4.25: Prediction table for VSCC

	B	M
B	303	54
M	188	24

Table 4.26: Prediction table for clustvarsel

	B	M
B	347	10
M	35	177

Table 4.27: ARI values, number of components, number of selected variables, and time from VSCC and clustvarsel analyses of the Wisconsin Breast Cancer data set.

	ARI	error	G	No. Variables	Time (secs)
VSCC	0.5254	13.71%	2	12	861.08
clustvarsel	0.7060	7.91%	2	6	936.61

Chapter 5

Discussion

This chapter discusses results from all data sets, special observations from Chapter 4, and comparisons for the summary table (Table 5.1). VSCC outperforms for the first 4 data sets and clustvarsel performs better for the other data sets. For the "Banknote" data set, both VSCC and clustvarsel reduce the variable numbers to 4 out of 6 but neither of them gets the correct number of clusters. When using the pre-determined cluster number of 2, the two methods have the same prediction tables that have the same ARI values and mis-clustering rates. The mean performance time for VSCC is much faster than for clustvarsel. Both methods reduce the number of variables for the "Coffee" data set. VSCC reduces the number of variables to 2 (total variable number is 12) which is less than clustvarsel which reduces the number of variables to 5. VSCC has a higher ARI value since VSCC has the correct cluster number but clustvarsel does not get the correct number. For the "Glass" data set, VSCC does not reduce the dimensionality of the data set and clustvarsel reduces the number of variables to 2 (out of 9). Both of the

methods do not get the correct number of clusters but VSCC performs a little bit better than clustvarsel with regard to the ARI value and the running time. The two methods have correct cluster numbers for the "Wine Recognition" data set. Clustvarsel selects 5 variables out of 13 and VSCC does not reduce the variable numbers, while VSCC has a higher ARI value than clustvarsel's ARI value. For the "Iris" data set, clustvarsel has the correct cluster number but VSCC does not, and clustvarsel also reduces the variables to a lower dimension. Regarding the ARI value and mis-clustering rate, clustvarsel performs better than VSCC. For the "Italian Olive Oil" data set, instead of clustering into 3 regions, this study clusters samples into 9 areas that belong to the 3 regions. Both of the methods have the correct cluster numbers and neither of the methods performs variable reduction. These probably are the reasons that the two methods end up with the same ARI values and mis-clustering rates, except that VSCC has a lower mean running time than clustvarsel. The "Leptograpsus Crab" data set originally has only two types, which are "Blue" and "Orange". After observing the clustering results, it is more reasonable to cluster the data set into 4 clusters that are "BM" (blue and male), "OM" (orange and male), "BF" (blue and female), and "OF" (orange and female). Both VSCC and clustvarsel have the correct cluster numbers and VSCC uses a fewer number of variables to cluster the samples. Clustvarsel has a higher ARI value and lower mis-clustering rate. For the "Wheat Kernels" data set, only clustvarsel has the correct cluster number which helps it get a higher ARI value and lower mis-clustering rate than VSCC. Clustvarsel chooses a fewer number of variables than VSCC for clustering the data set. The "Wisconsin Breast

Cancer" data set has 30 variables in total, VSCC uses 12 variables to cluster the samples and clustvarsel uses only 6 variables which leads clustvarsel outperforms VSCC in terms of the ARI value and the mis-clustering rate. Clustvarsel has a similar mean running time as VSCC for this data set.

Based on the discussions above, VSCC sometimes has problems selecting the correct cluster number. When VSCC gets the cluster number correct it performs well on ARI value and mis-clustering rate. Clustvarsel outperforms VSCC when it gets the cluster number correct in situations where VSCC gets the wrong cluster number. If the cluster number is pre-known then VSCC will work well on clustering the data set. Most of the time, clustvarsel performs well on dimension reduction. Consequently, if the data needs a lot of variable reduction clustvarsel may be the more appropriate technique.

The following table (Table 5.1) summarizes the ARI values and mis-clustering rates (error) for the VSCC and clustvarsel methods. The last column indicates the number of pairs of high correlation variables by assuming the high correlation value is 0.75. The table is arranged in an ascending order of the number of high correlation variables. Table 5.1 shows the "Banknote", "Glass Identification", "Coffee", , and "Wine Recognition" data sets have 0, 1, 2, and 2 pairs of high correlation variables respectively, and they have higher ARI values and lower mis-clustering errors for VSCC than clustvarsel. On the other hand, the "Iris", "Leptograpsus Crabs", "Wheat Kernel", and "Wisconsin Breast Cancer" data have 3, 10, 11, 48 pairs of high correlation variables respectively. These data have a higher ARI and lower mis-clustering error for clustvarsel than VSCC. The "Wheat Kernel" and

"Crab" data sets have variables with extreme high positive correlations and the clustvarsel method performs best here. There is an exception, the "Italian Olive Oil" data has the same ARI values and mis-clustering errors for the two methods with 4 pairs of high correlation variables but VSCC has shorter performance time than clustvarsel. Based on the observations, when only a few pairs of variables have high correlations, VSCC performs better than clustvarsel. When many pairs of variables are highly correlated then clustvarsel performs better than VSCC.

Table 5.1: Summary of ARI value, mis-clustering error, number of pairs with high correlation for all data sets.

	ARI		error		High corr
	VSCC	clustvarsel	VSCC	clustvarsel	
Banknote	0.8603	0.6907	43.00%	60.00%	0
Glass	0.1470	0.1465	69.16%	65.89%	1
Coffee	1	0.3732	0%	51.16%	2
Wine	0.9297	0.7828	2.25%	7.30%	2
Iris	0.5681	0.7196	33.33%	11.33%	3
Italian Olive Oil	0.6586	0.6586	31.29%	31.29%	4
Leptograpsus Crabs	0.8052	0.8291	8%	7%	10
Wheat Kernel	0.5075	0.8520	38.10%	5.24%	11
Breast Cancer	0.5254	0.7060	13.71%	7.91%	48

The two variable selection methods, VSCC and clustvarsel, are wrapper methods that use algorithms to select variables and the selection criterion depends on

the learning algorithm. For future study, there are other variable selection methods that can be considered, such as hierarchical sparse clustering, sparse k-means clustering, and Laplacian Score. It is possible to compare the results from different methods and find if there are methods that outperform VSCC or clustvarsel. The "Glass Identification" data set has a lower ARI value and a high mis-clustering error for both VSCC and clustvarsel, and the "Italian Olive Oil" data set have the same ARI values and mis-clustering rate for both methods. They are ideal data sets to use to study the application of other variable reduction techniques, with the ultimate goal being to improve the clustering accuracy.

Chapter 6

Conclusion

As the data dimensionality and complexity increases, it is costly and time consuming to use all the variables to cluster the data sets. Also, using a large number of variables for model-based clustering tends to over-parametrize the results. This study focus on finding efficient dimension reduction techniques for model-based clustering on high-dimensional data sets, and concluding which type of data set each technique might be better suited for by comparing the ARI values, the mis-clustering rates, and the performance time.

Two variable selection methods were compared in this study, VSCC and clustvarsel. Both of them are wrapper methods that use algorithms to select variables and the selection criterion depends on the learning algorithm. Based on the discussions in Chapter 5, it is possible to summarize that when VSCC gets the cluster number correct it performs well on the ARI values and mis-clustering error. Clustvarsel outperforms VSCC when it gets the cluster number correct in situations where VSCC gets the wrong cluster number. Also, clustvarsel performs

well on dimension reduction. In addition, when only a few pairs of variables have high correlations, VSCC performs better than clustvarsel. When many pairs of variables are highly correlated, clustvarsel performs better than VSCC with regard to the ARI values and the mis-clustering rates. VSCC also has shorter mean performance time than clustvarsel on all of the data sets.

References

Andrews, J. L., & McNicholas, P. D. (2014). Variable selection for clustering and classification. *Journal of Classification*, *31*(2), 136–153.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Bartolucci, F., Montanari, G. E., & Pandolfi, S. (2016). Item selection by latent class-based methods: an application to nursing home evaluation. *Advances in Data Analysis and Classification*, *10*(2), 245–262.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). The new s language. wadsworth & brooks/cole. *Computer Science Series, Pacific Grove, CA*.

Bhattacharya, S., & McNicholas, P. D. (2014). A lasso-penalized bic for mixture model selection. *Advances in Data Analysis and Classification*, *8*(1), 45–61.

Blake, C. L., & Merz, C. J. (1998). *Uci repository of machine learning databases, 1998.*

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for bart: an application to gene regulation. *The Annals of Applied Statistics*, 1750–1781.

Bontemps, D., & Toussile, W. (2013). Clustering and variable selection for categorical multivariate data. *Electronic Journal of Statistics*, *7*, 2344–2371.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, *71*, 52–78.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., & Żak, S. (2010). Complete gradient clustering algorithm for features analysis of x-ray images. In *Information technologies in biomedicine* (pp. 15–24). Springer.

Clogg, C. C. (1988). Latent class models for measuring. In *Latent trait and latent class models* (pp. 173–205). Springer.

Dean, N., & Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, *62*(1), 11.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, *6*(4), 45.

Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. *Journal of machine learning research*, *5*(Aug), 845–889.

Evett, I. W., & Ernest, J. S. (1987). Rule induction in forensic science. central research establishment. home office forensic science service. aldermaston. *Reading, Berkshire RG7 4PN*.

Flury, B. (1988). *Multivariate statistics: a practical approach*. Chapman & Hall, Ltd.

Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, *12*, 18–65.

Fop, M., Smart, K. M., & Murphy, T. B. (2017). Variable selection for latent class analysis with application to low back pain diagnosis. *The Annals of Applied Statistics*, 2080–2110.

Forina, M., Armanino, C., Lanteri, S., & Tiscornia, E. (1983). Classification of olive oils from their fatty acid composition. In *Food research and data analysis: proceedings from the iufost symposium, september 20-23, 1982, oslo, norway/edited by h. martens and h. russwurm, jr.*

Forina, M., & Tiscornia, E. (1982). Pattern-recognition methods in the prediction of italian olive oil origin by their fatty-acid content. *Annali di Chimica*, *72*(3-4), 143–155.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of classification*, *5*(2), 205–228.

Fraley, C., & Raftery, A. E. (2006). *Mclust version 3 for r: Normal mixture modeling and model-based clustering* (Tech. Rep.). Citeseer.

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). *mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation* (Tech. Rep.). Technical report.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. Springer Science & Business Media.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American journal of public health*, *79*(3), 340–349.

Houseman, E. A., Coull, B. A., & Betensky, R. A. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics*, *62*(4), 1062–1070.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

Jeffrey L. Andrews, P. D. M. (2013). vscc: Variable selection for clustering and classification [Computer software manual]. (R package version 0.2)

Johnstone, I. M., & Titterington, D. M. (2009). *Statistical challenges of high-dimensional data.* The Royal Society Publishing.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Law, M. H., Figueiredo, M. A., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, *26*(9), 1154–1166.

Lawrence, C., Liu, J., Palumbo, M., & Zhang, J. (2003). Bayesian clustering with variable and transformation selections. *Bayesian statistics*, *7*, 249–275.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *Nsf-cbms regional conference series in probability and statistics* (pp. i–163).

Liu, H., & Motoda, H. (2007). *Computational methods of feature selection*. CRC Press.

Marbac, M., & Sedki, M. (2017a). Variable selection for mixed data clustering: a model-based approach. *arXiv preprint arXiv:1703.02293*.

Marbac, M., & Sedki, M. (2017b). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, *27*(4), 1049–1063.

Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009a). Variable selection for clustering with gaussian mixture models. *Biometrics*, *65*(3), 701–709.

Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009b). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, *53*(11), 3872–3882.

Maugis-Rabusseau, C., Martin-Magniette, M.-L., & Pelletier, S. (2012). Selvarclustmv: Variable selection approach in model-based clustering allowing for missing values. *Journal de la société française de statistique*, *153*(2), 21–36.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 38). M. Dekker New York.

McNicholas, P. D. (2016). *Mixture model-based classification*. Chapman and Hall/CRC.

Mengersen, K. L., Robert, C., & Titterington, M. (2011). *Mixtures: estimation and applications* (Vol. 896). John Wiley & Sons.

Neath, A. A., & Cavanaugh, J. E. (2012). The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.

Nia, V. P., & Davison, A. C. (2012). High-dimensional bayesian clustering with variable selection: The r package bclust. *Journal of Statistical Software*, *47*(ARTICLE), 1–22.

Pan, W., & Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of machine learning research*, *8*(5).

Peel, D., & MacLahlan, G. (2000). Finite mixture models. *John & Sons*.

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, *101*(473), 168–178.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846–850.

Ripley, B. D. (2002). *Modern applied statistics with s*. springer.

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

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, *8*(1), 289.

Scrucca, L., & Raftery, A. E. (2018). clustvarsel: a package implementing variable selection for gaussian model-based clustering in r. *Journal of Statistical Software*, *84*.

Sedki, M., Celeux, G., Maugis-Rabusseau, C., Sedki, M. M., Rcpp, I., & Rcpp, L. (2014). Package 'selvarmix'.

Silvestre, C., Cardoso, M. G., & Figueiredo, M. (2015). Feature selection for clustering categorical data with an embedded modelling approach. *Expert systems*, *32*(3), 444–453.

Streuli, H. (1973). Der heutige stand der kaffeechemie. In *Assic, 6e. colloque, bogota* (Vol. 61).

Swayne, D. F., Cook, D., Buja, A., Lang, D. T., Wickham, H., & Lawrence, M. (2006). Ggobi manual. *Sourced from www. ggobi. org/docs/manual. pdf*.

Tadesse, M. G., Sha, N., & Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, *100*(470), 602–617.

Toussile, W., & Gassiat, E. (2009). Variable selection in model-based clustering using multilocus genotype data. *Advances in Data Analysis and Classification*, *3*(2), 109–134.

Witten, D. M., Tibshirani, R., & Witten, M. D. (2013). Package 'sparcl'.

Wolfe, J. H. (1963). *Object cluster analysis of social areas* (Unpublished doctoral dissertation). University of California.

Wu, B. (2013). Sparse cluster analysis of large-scale discrete variables with application to single nucleotide polymorphism data. *Journal of applied statistics*, *40*(2), 358–367.