Variable Selection for Skewed Clustering and Classification

VARIABLE SELECTION FOR SKEWED CLUSTERING AND CLASSIFICATION

BY

MACKENZIE R. NEAL, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2022)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario, Canada

TITLE:	Variable Selection for Skewed Clustering and Classification
AUTHOR:	Mackenzie R. Neal B.Sc., (Mathematical Science) University of Guelph, Guelph, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	ix, 39

To my parents, Sandy and Jeff.

Abstract

As datasets from virtually all fields of endeavour continue to grow in size and complexity, the curse of dimensionality cannot be overlooked. Researchers in model-based clustering have recognized the need for effective dimension reduction techniques; as a result, many such algorithms exist to date. These algorithms, however, are often specific to Gaussian clustering problems and break down in the presence of skewness. We present a novel skewed variable selection algorithm that utilizes the Manly transformation mixture model to select variables based on their ability to separate clusters. We compare our approach with other asymmetric and normal variable selection methods using simulated and real-world datasets. We find that the proposed algorithm is suitable for dimension reduction in the presence of skewness.

Acknowledgements

First and foremost, I would like the thank my supervisor, Dr. Paul McNicholas. Your guidance, support, and encouraging words have been fundamental in completing this thesis. Thank you for continuously pushing me beyond what I believed I was capable of; I have grown personally and professionally because of it.

I would like to extend my gratitude toward the faculty and staff within the Department of Mathematics and Statistics at McMaster University. Firstly, to Dr. Noah Forman and Dr. Traian Pirvu for their roles on my examination committee. Secondly, to all my professors, I am grateful to have learned from each of you. I would also like to thank the department staff for always being incredibly responsive and helpful.

To my friends and office mates at McMaster, thank you for answering my countless questions and for making me feel welcome. Finally, thank you to my parents for showing me what hard work looks like, for always answering my calls - no matter how inconvenient it was for you, and for an abundance of love and support. I could not have done this without you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Variable selection refers to the process by which informative variables are retained and uninformative variables are removed. Eliminating uninformative variables can improve both model fitting and model interpretability. As such, much research has been conducted on variable selection across statistical domains. One such domain is that of model-based clustering and classification. The need for dimension reduction is evident for clustering and classification problems as noisy data can hide key features, such as groupings. We know that dimension reduction should happen in tandem with data clustering rather than before clustering (Steinley and Brusco, 2011; Bouveyron and Brunet-Saumard, 2014). As such, variable selection methods that are embedded into clustering and classification algorithms are essential. Many such algorithms exist for Gaussian clustering algorithms; the same cannot be said for skewed clustering methods.

In this paper, we study the effect that skewness has on existing variable selection algorithms for classification and clustering and introduce a skewed extension to the popular variable selection method VSCC (Andrews and McNicholas, 2014), which is available as the `vscc` package (Andrews and McNicholas, 2013) for R (R Core Team, 2022). We compare this extension to a skewed extension of the clustvarsel algorithm (Wallace *et al.*, 2018), using both real data and simulated data in Chapter 4. In Chapter 2, we briefly discuss existing variable selection algorithms and methods for skewed model-based clustering. We introduce our skewed extension to the VSCC algorithm and discuss data analysis details in Chapter 3. Lastly, in Chapter 5, we include a discussion of results and provide suggestions for future work.

Chapter 2

Background

2.1 Finite Mixture Models

Finite mixture models arise from the assumption that a population contains sub-populations that can be modelled by a finite number of densities. Thus, these models lend themselves to clustering and classification problems quite nicely. A random vector $\mathbf{X}$ belongs to finite mixture model if, for all $\mathbf{x} \subset \mathbf{X}$ we can write the density as

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

where $\pi_g > 0$ are the mixing proportions such that $\sum_{g=1}^{G} \pi_g = 1$ and $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ are the component densities. Most commonly, these component densities are taken to be multivariate Gaussian resulting in the following finite mixture model density,

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)\right\}.$$

However, in real application it is uncommon for data to be fully Gaussian. Thus various asymmetric mixture models have been developed to aid in clustering and classification when skewness is present, as discussed in Section 2.2. More comprehensive details on finite mixture models can be found in Everitt and Hand (1981), Titterington *et al.* (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006).

2.2 Skewed Mixture Models

There are two schools of thought when it comes to dealing with skewness. The first accounts for skewness directly with the use of flexible, asymmetric distributions. These include skew-symmetric distributions such as the skew-normal with density (Pyne *et al.*, 2009):

$$f(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi_1(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu});\mathbf{0},\mathbf{1}-\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}),$$

where ϕ_p and Φ_p are the pdf and cdf, respectively, of the standard multivariate normal; $\boldsymbol{\Sigma}$ is the covariance matrix; $\boldsymbol{\delta}$ is the vector of skewness parameters; and $\boldsymbol{\mu}$ is the location parameter vector. Other common asymmetric distributions include the family of generalized hyperbolic distributions (Browne and McNicholas, 2015) such as the normal inverse Gaussian (Karlis and Santourian, 2009), variance-gamma (McNicholas *et al.*, 2014), and the shifted asymmetric Laplace (Franczak *et al.*, 2014) distributions. These distributions are also known as normal variance-mean mixtures. A p-dimensional random vector $\mathbf{X}$ is a normal variance-mean mixture if its density can be written as

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \int_0^\infty \phi_p(\mathbf{x}|\boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Sigma})h(y|\boldsymbol{\theta})dy,$$

where $\phi_p(\mathbf{x}|\boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Sigma})$ is the density of a p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu} + y\boldsymbol{\alpha}$ and covariance $y\boldsymbol{\Sigma}$ and $h(y|\boldsymbol{\theta})$ is a density function for an asymmetric random variable $Y > 0$ (Barndorff-Nielsen *et al.*, 1982). In Section 4.2, we generate data from a mixture of multivariate variance gamma distributions to compare variable selection methods in the presence of skewness. Data from a multivariate variance-gamma distribution can be generated via

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{U},$$

where $Y \sim \text{gamma}(\lambda, \psi/2)$ and $\mathbf{U} \sim N_p(0, \boldsymbol{\Sigma})$ to result in $\mathbf{X} \sim V_p(\lambda, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ (McNicholas *et al.*, 2014).

The other school of thought for dealing with skewness utilizes transformations to near-normality. Two transformation mixture models exist; the first is a t-mixture model with a Box-Cox transformation (Lo and Gottardo, 2012). This model, however, would suffer from the shortcomings of the Box-Cox transformation, primarily its inability to handle left skew (Box and Cox, 1964). Additionally, the Box-Cox t-mixture assumes a global transformation parameter, thus, transformations do not vary by variables and components (Lo and Gottardo, 2012). The second transformation mixture model is a normal mixture model with a Manly transform (Zhu and Melnykov, 2018a). The Manly can handle both left and right skewed data and can be applied to any real number. We will use the latter transformation, given by

$$T(\mathbf{x}|\lambda) = \begin{cases} \frac{\exp\{\lambda \mathbf{x}\}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \mathbf{x}, & \text{otherwise.} \end{cases}$$

By applying the back transform of the Manly, one will arrive at the following transformation-based density:

$$f_T(\mathbf{x}|\boldsymbol{\vartheta}) = \phi(T(\mathbf{x}|\boldsymbol{\Lambda}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) J_T(\mathbf{x}|\boldsymbol{\Lambda}).$$

where $\mathbf{x}$ is the original p-dimensional data vector; $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, ..., \lambda_p)$ is the transformation vector; $\boldsymbol{\mu}$ is the location parameter vector and $\boldsymbol{\Sigma}$ is the covariance matrix, and the Jacobian of the back transformation can be written as

$$J_T(\mathbf{x}|\boldsymbol{\Lambda}) = \exp\{\boldsymbol{\Lambda}'\mathbf{x}\},$$

and Zhu and Melnykov (2018a) have utilized this back transformation to obtain a skewed finite mixture model.

This mixture model contains transformation parameters for each variable-cluster combination. As such, by incorporating the Manly into a model one must introduce $G \times p$ additional transformation parameters, potentially resulting in over-parameterization the model. To overcome this, Zhu and Melnykov (2018a) recognized that it is unlikely for all variables to need to be transformed in all components. Thus to avoid over-parameterization, unnecessary transformation parameters are determined and zeroed out via a backwards or forwards selection process.

Forwards selection begins with a fully Gaussian mixture model (GMM). The GMM is then compared to $G \times p$ models each with one non-zero transformation parameter.

The value of this non-zero transformation parameter is selected based on the simplex method, where the conditional expectation of the complete-data log-likelihood is maximized with respect to the skewness parameter in question. For each resulting model, BIC is obtained as follows:

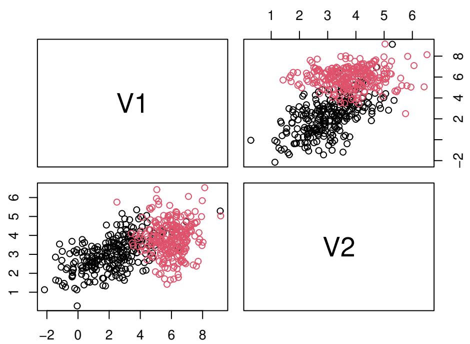
$$\text{BIC} = p\log(n) - 2\log(\hat{L}),$$

where $\hat{L}$ is the maximized likelihood estimate (Schwarz, 1978). Among the $G \times n$ candidates, we select the model that minimizes BIC. The algorithm continues until there are no improvements to BIC, where parameters from the previous step are used for initializations of the next step.

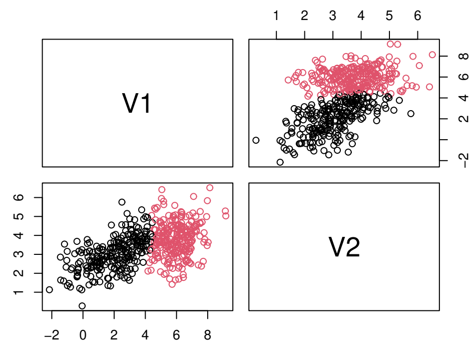
Backwards selection begins with a fully skewed Manly mixture model, iteratively one transformation parameter is zeroed out and BIC is obtained and compared. Again, this process is continued until no more improvements to BIC are observed. We utilize the work of Zhu and Melnykov (2018a) on the Manly mixture to extend the VSCC algorithm into the skewed space, this extension is detailed in Section 3.1.

2.3 Variable Selection

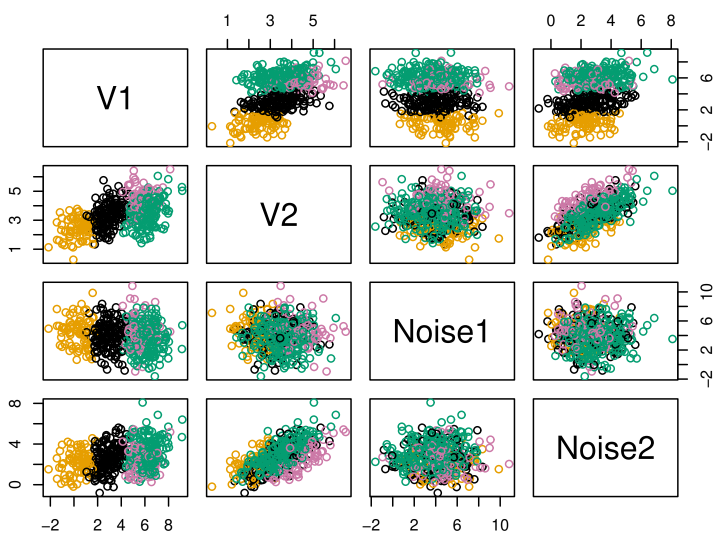
The need for dimension reduction algorithms for model-based clustering is evidenced in Figure 2.1, where we simulate data from a two-component, two-dimensional GMM and fit a GMM to this data before and after the addition of two noise variables. The first noise variable is random noise generated from a normal distribution with mean four and standard deviation two. The second noise variable is correlated to the second clustering variable, $\text{Noise}_2 = 0.8 * V2 + 0.2 * Z$ where $Z \sim N(0, 5)$. We see that with the addition of just two noise variables, the clustering results begin to break down.



(a) True clusters from two-component GMM



(b) Clustering results from GMM fit to simulated data in (a)



(c) Clustering results from GMM fit to simulated data in (a) plus two noisy variables.

Figure 2.1: Clustering results from GMM on noisy data

One type of dimension reduction method that could be used to overcome the poor clustering performance seen in Figure 2.1 is variable selection. Variable selection is the selection of important variables and the de-selection of unimportant variables. Many such variable selection algorithms for clustering and classification exist to date; summaries of said algorithms can be found in various papers (Steinley and Brusco, 2008; Adams and Beling, 2019; Fop and Murphy, 2018). The two most commonly used algorithms, due to both performance and availability, are `clustvarsel` (Scrucca and Raftery, 2018; Raftery and Dean, 2006; Maugis *et al.*, 2009) and `vscc` (Andrews and McNicholas, 2013).

2.3.1 `clustvarsel`

The `clustvarsel` algorithm makes use of three sets of variables to perform variable selection. The first is the set containing selected variables X_{clust}, the second is the variable under consideration for inclusion or exclusion X_i, and the third contains all remaining variables X_{other}. The Bayes factor is used to compare two models essential for variable selection. The first model assumes X_i is unimportant for clustering but is related to the set, or a subset, of the clustering variables through linear regression. The integrated likelihood for this model, denoted by $f_1(X_{\text{clust}}, X_i|M_1)$ where M_1 is the selected G-component Gaussian mixture model, can be decomposed into the following

$$f_1(X_{\text{clust}}, X_i|M_1) = f_{\text{reg}}(X_{\text{clust}}, X_i)f_{\text{clust}}(X_{\text{clust}}|M_1),$$

where $f_{\text{reg}}(X_i|X_{\text{clust}})$ is the regression of X_i onto the set, or a subset, of the clustering variables. This subset is selected through stepwise regression, wherein variables from the clustering set are selected if they aid in the prediction of X_i. Model one is compared to a second model where X_i is important to clustering and thus the integrated likelihood becomes

$$f_2(X_{\text{clust}}, X_i|M_2) = f_{\text{clust}}(X_{\text{clust}}, X_i|M_2).$$

The Bayes factor can then be determined as the following

$$\mathrm{B}_{12} = \frac{f_1(X_{\text{clust}}, X_i|M_1)}{f_2(X_{\text{clust}}, X_i|M_2)}.$$

As integrated likelihoods are difficult to compute, $-2\log \mathrm{B}_{12}$ is approximated by $\mathrm{BIC}_{\text{diff}}$, defined as

$$\begin{aligned}\mathrm{BIC}_{\text{diff}} &= \mathrm{BIC}_{\text{clust}}(X_{\text{clust}}, X_i) - \mathrm{BIC}_{\text{not clust}}(X_{\text{clust}}, X_i)\\ &= \mathrm{BIC}_{\text{clust}}(X_{\text{clust}}, X_i) - \mathrm{BIC}_{\text{clust}}(X_{\text{clust}}) - \mathrm{BIC}_{\text{reg}}(X_i|X_{\text{clust}}),\end{aligned}$$

where $\mathrm{BIC} = 2\log(\hat{L}) - p\log(n)$, the negative of the BIC formula seen in Schwarz (1978). Thus, a positive $\mathrm{BIC}_{\text{diff}}$ corresponds to a small Bayes factor, which would suggest that we should cluster on both X_i and X_{clust}. The `clustvarsel` algorithm iterates between inclusion and exclusion steps, where one by one the variables not in X_{clust} are considered for inclusion and variables in X_{clust} are considered for exclusion. Variables that maximize $\mathrm{BIC}_{\text{diff}}$ are included and variables that minimize $\mathrm{BIC}_{\text{diff}}$ are removed. As dimensions increase this algorithm becomes increasingly slow due to its step-wise nature. Additionally, `clustvarsel` will perform poorly in the presence of skewed clusters due to its reliance on Gaussian mixture models.

Wallace *et al.* (2018) extend `clustvarsel` into the skewed space with the use of the multivariate skew-normal distribution (Pyne *et al.*, 2009). The multivariate skew-normal (MSN) is known to be a restrictive asymmetric distribution and normal-like in the tails, thus making it less robust to outlying observations. Regardless, Wallace *et al.* (2018) select the MSN for the skewed extension of `clustvarsel` due to its computational efficiency, robustness to starting values, and the availability of both regression and mixture model estimation tools, as each are needed in the variable selection laid out by Maugis *et al.* (2009) and extended by Scrucca and Raftery (2018) to implement `clustvarsel`.

2.3.2 `vscc`

The `vscc` algorithm selects variables based on minimization of within-cluster variance and maximization of between-cluster variance. These goals can be met simultaneously when the data is scaled prior to implementation of the algorithm. The `vscc` algorithm tends to be much faster than `clustvarsel` as we perform model fitting on only the original and the final selected variables, rather than at every inclusion/exclusion step. The algorithm begins by calculation of the within-group variance for each variable. The variable that minimizes within-group variance the most is automatically selected into the clustering set. From there, variables are selected into the clustering set based on their ability to separate clusters and their correlation to the set of selected variables. A moving selection criterion is used to do so. This criterion begins with a linear relationship between within-group variance W_j and correlation ρ_{jr} and moves to a quintic relationship. Variable j is selected into the clustering set V_i if for all

$r \in V_i$ the following criteria holds

$$|\rho_{jr}| < 1 - W_j^i.$$

As i increases, the correlation criteria is loosened to allow more correlation between the selected variables. A graphical representation of this relationship, similar to Figure 1 found in Andrews and McNicholas (2013), can be found below in Figure 2.2. The `vscc` algorithm tests five exponent values $i = 1, 2, ..., 5$, resulting in five potential subsets of selected variables. Model-based clustering is carried out on each subset and the final subset is selected based on minimization of clustering uncertainty. With the use of the soft classification matrix Andrews and McNicholas (2014) define uncertainty to be

$$n - \sum_{i=1}^{n} \max_g (\hat{z}_{ig}),$$

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

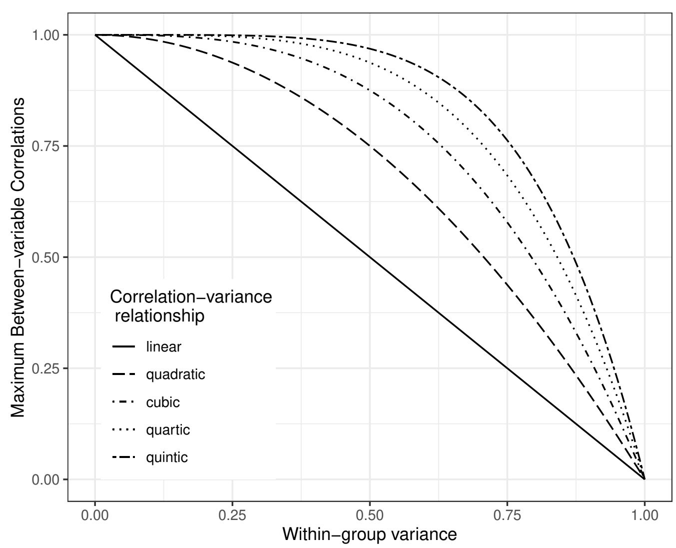


Figure 2.2: Correlation-variance relationship for selection criteria.

The `vscc` algorithm is computationally efficient and performs well on Gaussian clusters. However, as variables are selected based on the minimization of within-cluster variance this method would suffer substantially when applied to skewed clusters. As such, Chapter 3 discusses how this algorithm could be extended to skewed data and Chapter 4 compares said extension to the previously discussed algorithms.

Chapter 3

Methodology

3.1 Algorithm

We must transform the data to near-normality for minimization of within-cluster variance to be used as a variable selection criterion for skewed clustering/classification problems. Thus, we propose an extension to `vscc` where a Manly mixture is fit to the data. The transformation parameters are then obtained from the fitted model and applied to the data prior to conducting the variable selection laid out in `vscc`. The skewed clustering extension is detailed below in Algorithm 1 where $g = 1, ..., G$ refers to the cluster number, $i = 1, ..., n$ is the index of points, $j = 1, ..., p$ is the variable number, and $\hat{z}_{ig}$ is the group membership obtained from clustering,

$$\hat{z}_{ig} = \begin{cases} 1, & \text{if observation } \mathbf{x}_i \text{ belongs to group } g \\ 0, & \text{otherwise.} \end{cases}$$

Algorithm 1 VSCC Manly

1: Perform model-based clustering, fitting either a full Manly mixture or a Manly mixture with transformation parameter selection.
2: Use $\hat{z}_{ig}$ as initial group memberships.
3: Transform data according to the following,

$$\mathbf{Y}_g = \left(\frac{e^{\lambda_{1g}\mathbf{x}_1} - 1}{\lambda_{1g}}, \cdots, \frac{e^{\lambda_{pg}\mathbf{x}_p} - 1}{\lambda_{pg}} \right)$$

4: Scale transformed variables.
5: Calculate within-group variance for each variable,

$$\hat{W}_j = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig}(y_{ji} - \hat{\mu}_{jg})^2}{n}$$

6: Sort $\hat{W}_j$ in ascending order.
7: **for** `i in` $1:5$ **do**
8:     $\hat{W}_1$ is automatically selected into $V_{(i)}$, j=2
9:     **if** $|\rho_{jr}| < 1 - \hat{W}_j^i$ for all r in $V_{(i)}$ **then**
10:         Variable s=j is placed in $V_{(i)}$
11:     **else**
12:         Variable is not placed in $V_{(i)}$
13:     **if** $j < p$ **then**
14:         set j=j+1 and return to line 9
15: Perform model-based clustering with a Manly mixture on all five variable subsets.
16: Select $V_{(i)}$ such that $n - \sum_{i=1}^{n} max_g(\hat{z}_{ig})$ is minimized.

Algorithm 1 details the skewed extension of `vscc` for clustering problems. For this method to be applied to classification problems transformation parameters and true group memberships, z_{ig}, would need to be supplied in replacement of lines one and two in Algorithm 1. Transformation parameters can be determined by maximizing the expectation of the complete-data log likelihood with respect to the transformation parameters.

We note that studies on traditional skewed methods vs. transformation methods have found that no one type of method for handling skewness outperforms the other (Gallaugher *et al.*, 2020); thus, the use of a transformation-based mixture model is an appropriate choice for dealing with skewness. Additionally, we select a transformation-based mixture model for extending this algorithm into the skewed space as a direct, asymmetric distribution would not allow for transformation of clusters.

3.2 Initializations

Both `vscc` and `clustvarsel` are dependent on the R package `mclust` (Scrucca *et al.*, 2016); as such they use the `mclust` defaults for model fitting. For model initialization this is hierarchical clustering. As a result, the same model and selected variables will be obtained every time the algorithm runs on a given dataset. Both k-means and hierarchical are possible initialization schemes for `vscc-manly`. Due to the randomization of initial centres, k-means starts can result in different final models and selected variables. To control for this behaviour we run `vscc-manly` five times, once with a hierarchical start and four times with a k-means start. We then select the most common result; if multiple results are equally common, then the result that minimizes clustering uncertainty is selected. To remain consistent with Wallace *et al.* (2018), we run the `skewvarsel` algorithm five times with k-means starts, selecting the most common method that minimizes uncertainty.

3.3 Performance Assessment

Performance can be easily measured for simulated data as we know the clustering variables *a priori*. For real data, there are no true clustering variables; as a result, measuring performance becomes more difficult. We measure performance in three ways: adjusted Rand index (ARI), the number of clusters chosen, and visually with variable plots. As the dimension of the selected set increases, it becomes harder to assess performance using visuals. Regardless, one can still observe redundancy in the selected set, and thus, variable plots remain helpful even in such circumstances. We are operating in the clustering framework; however, true labels exist for all datasets tested. Therefore, ARI remains a valuable performance measure. Prior to ARI, the Rand index (RI) was used to compare partitions (Rand, 1971):

$$\mathrm{RI} = \frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}}.$$

ARI was proposed as to force the index to have expected value of zero under random assignment (Hubert and Arabie, 1985). The corrected index can be found below

$$\mathrm{ARI} = \frac{\mathrm{RI} - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}}.$$

Thus, ARI equals one when there is perfect agreement between partitions and is negative when the assignment is worse than random.

3.4 Model Fitting

All previously discussed methods will be tested on each dataset. To ensure fair comparison between `vscc-manly` and `skewvarsel`, we fit both a MSN mixture and a Manly mixture to the variables selected by `skewvarsel`. An MSN mixture is fitted to remain consistent with Wallace *et al.* (2018) and with the `skewvarsel` algorithm, as the BIC used for variable selection comes from a MSN mixture. The Manly mixture is fitted to help ensure that ARI performance is based on the variables selected and not the appropriateness of the distribution for the data in question.

Chapter 4

Analyses

4.1 Real Data Results

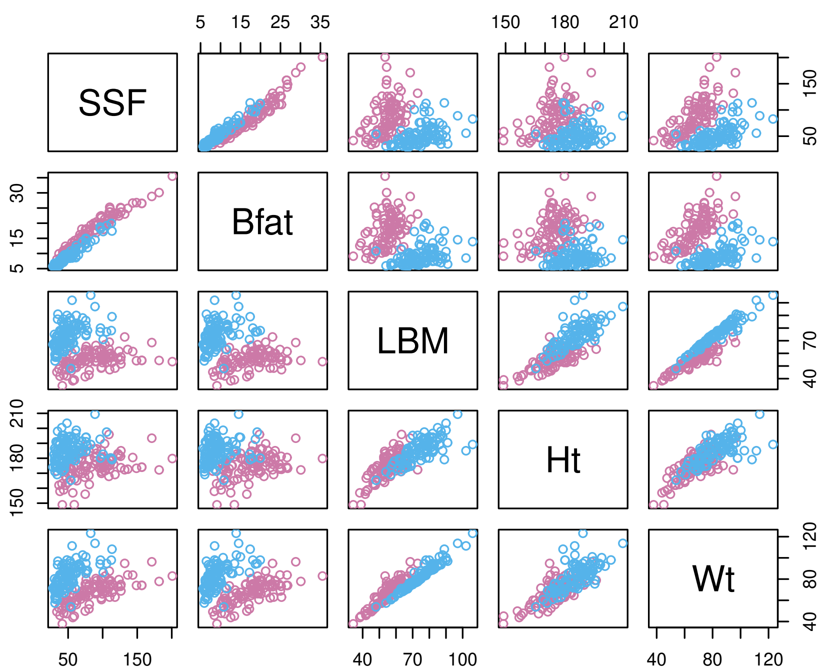
The `vscc`, `clustvarsel`, `vscc-manly`, and `skewvarsel` algorithms are compared on four datasets under a clustering framework. All methods will test $G = 1, ..., 9$ and data is standardized prior to running each method.

4.1.1 Australian Institute of Sport Data

The Australian Institute of Sport (AIS) dataset can be found in the `ManlyMix` package (Zhu and Melnykov, 2018b). This dataset contains 11 measurements on 202 individuals. Clustering results are compared to the sex column. From Table 4.1 we find that the `vscc-manly` algorithms perform the best in terms of G and ARI. More significantly, the `vscc-manly`-forwards algorithm reduces the dimensions more than all other methods tested. From Figure 4.1, we see that the variables selected by the `vscc-manly` algorithms clearly separate the true clusters. All other methods tested appear to be more susceptible to correlated variables, thus creating redundancy in the selected set. The `vscc-manly`-forwards and `vscc-manly`-backwards algorithms

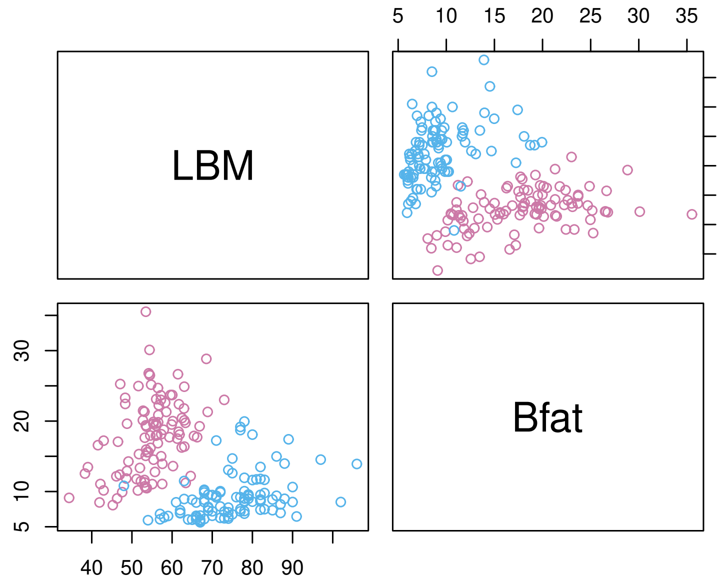
Table 4.1: Variables selection results for the AIS data.

Model	G	ARI	Variables
`vscc`	4	0.61	LBM, Bfat, SSF, Wt, Ht
`clustvarsel`	7	0.27	LBM, Bfat, Wt
`vscc-manly`-forward	2	0.94	LBM, Bfat
`vscc-manly`-backward	2	0.96	LBM, Bfat, Hg
`vscc-manly`-full	2	0.96	LBM, Bfat, Hg
`skewvarsel` + MSN	3	0.26	LBM, Bfat, SSF, Wt
`skewvarsel` + Manly forward	4	0.59	LBM, Bfat, SSF, Wt
`skewvarsel` + Manly backward	4	0.57	LBM, Bfat, SSF, Wt

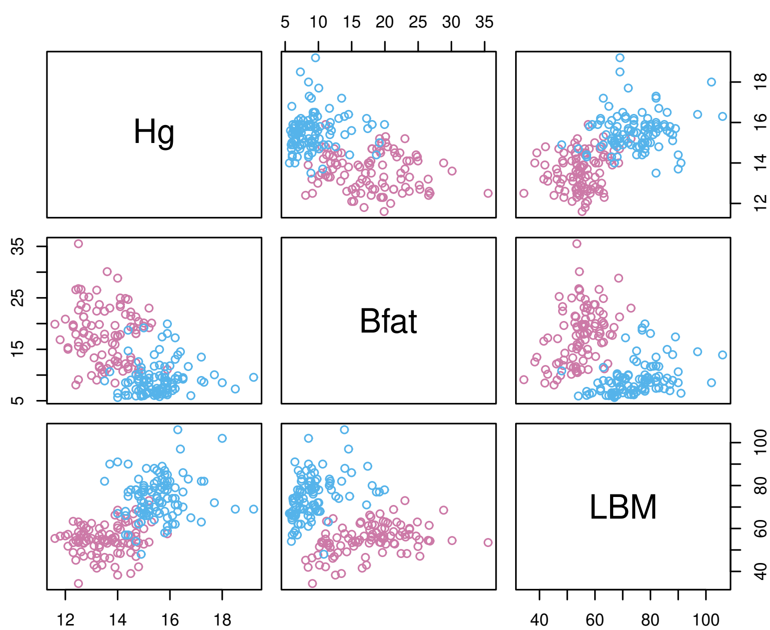


(a) Variables selected by `vscc`.

Figure 4.1: Plots of variables selected from AIS dataset.

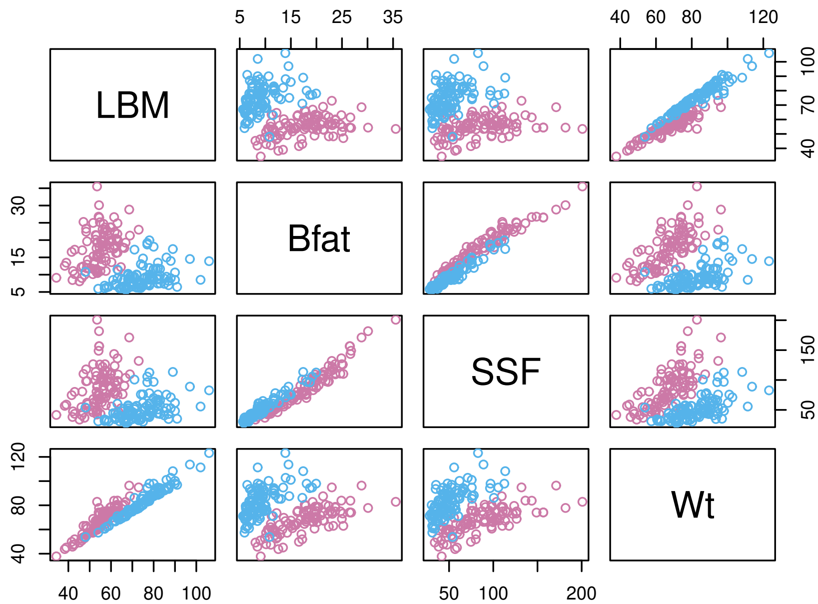


(b) Variables selected by `vscc-manly`-forwards.



(c) Variables selected by `vscc-manly`-backwards.

Figure 4.1: Plots of variables selected from AIS dataset.



(d) Variables selected by `skewvarsel`.

Figure 4.1: Plots of variables selected from AIS dataset.

resulted in the selection of a different final set of variables. Just as forwards and backwards step-wise regression can result in different results, forwards and backwards transformation parameter selection can result in different transformed spaces. As a result, it is unsurprising to see a difference in the set of selected variables between these methods.

4.1.2 Banknote Data

The banknote dataset comes from the mclust package (Scrucca *et al.*, 2016). There are six measurements, 200 observations, and two types of bills (genuine and counterfeit) of which clustering results are compared to.

Variable selection on the banknote dataset produces interesting results as no method significantly reduces dimensions (Table 4.2). This is surprising because, from Figure 4.2, it appears as though only two variables would be necessary for separating clusters. However, when a Manly mixture is fit to either the variables selected by the `skewvarsel` or the `vscc` algorithm, three clusters are found and ARI drops to 0.85. This suggests that although it may seem like the `vscc-manly` algorithm is selecting too many variables, the algorithm may be selecting the number of variables needed to ensure higher clustering performance.

Table 4.2: Variable selection results for the banknote data.

Model	G	ARI	Variables
`vscc`	3	0.86	Diagonal, Bottom, Top, Right
`clustvarsel`	4	0.67	Diagonal, Bottom, Top, Left, Length
`vscc-manly`-forward	2	0.98	Diagonal, Bottom, Top, Right, Left
`vscc-manly`-backward	2	0.98	Diagonal, Bottom, Top, Right, Left
`vscc-manly`-full	2	0.98	Diagonal, Bottom, Top, Right, Left
`skewvarsel` + MSN	4	0.69	Diagonal, Bottom, Top, Left
`skewvarsel` + Manly forward	3	0.85	Diagonal, Bottom, Top, Left
`skewvarsel` + Manly backward	3	0.85	Diagonal, Bottom, Top, Left

4.1.3 Italian Wine Data

The Italian wine dataset can be found in the `pgmm` package (McNicholas *et al.*, 2022). It contains 28 variables, 178 observations, and three types of wine of which clustering results are compared to.

In Table 4.3, we see that `vscc` performs the best while `vscc-manly` performs the worst, in terms of G and ARI. All methods appear to reduce dimensions approximately the same amount with key variables such as flavanoids and hue being selected nearly every time.

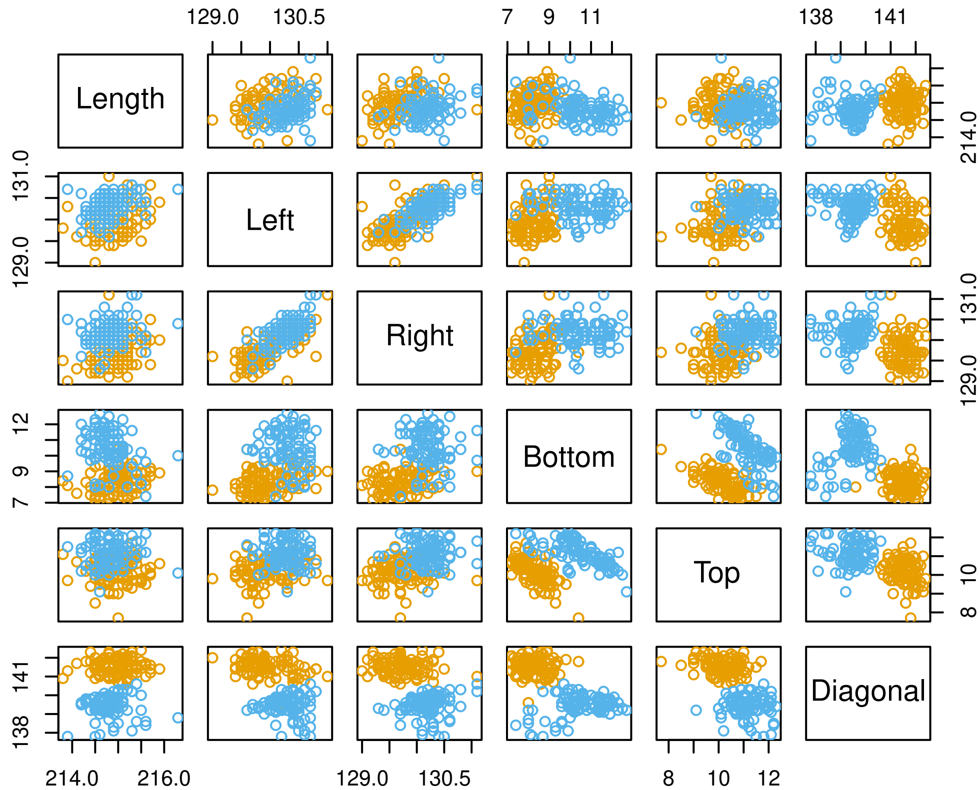


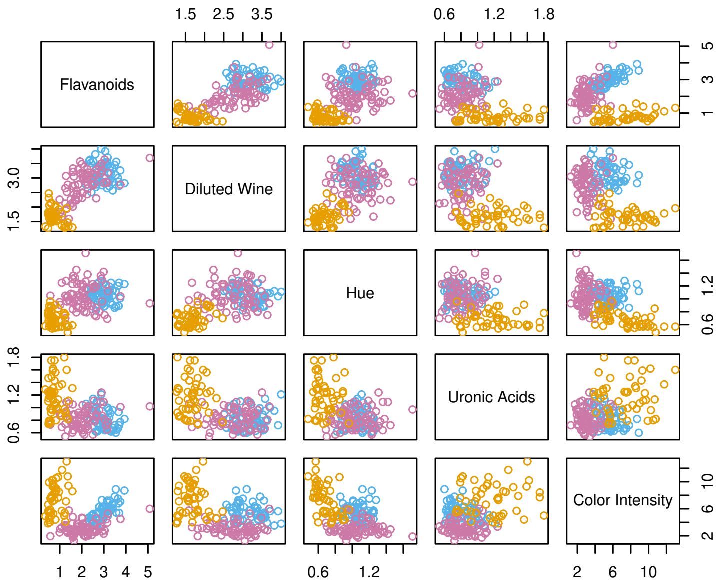
Figure 4.2: Variables in the banknote data.

This would suggest that it is not the minimization of within-cluster variance that is performing poorly on this dataset but rather the fit of the Manly mixture. This point is further emphasized when we look at the `skewvarsel` results in Table 4.3. When the MSN mixture is fit to the `skewvarsel` selected variables, a much higher ARI is obtained than when the backwards Manly mixture is fit to the same variables. These results suggest that the Manly may be more prone to combining Gaussian clusters to create skewed clusters. As the MSN distribution is normal-like in the tails, the

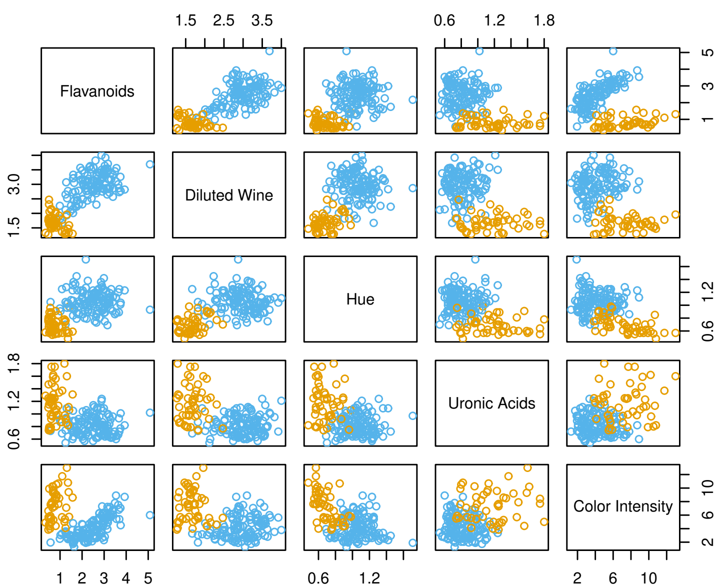
MSN may be less prone to the same behaviour. We illustrate this on simulated data from a three-component, two-dimensional GMM in Figure 4.4. This behaviour is also seen in the pairs plots of the Italian wine data selected by `vscc-manly`-backwards (Figure 4.3).

Table 4.3: Variable selection results from the Italian wine dataset.

Model	G	ARI	Variables
`vscc`	3	0.90	Flavanoids, Hue, OD280/OD315 Diluted Wine, Proline, Colour Intensity, Alcohol,Total Phenols
`clustvarsel`	5	0.67	Flavanoids, Proline, Colour Intensity, Uronic Acid Chloride, Malic Acid
`vscc-manly`-forward	2	0.43	Flavanoids, Hue, OD280/OD315 Diluted Wine, OD280/OD315 Flavanoids
`vscc-manly`-backward	2	0.49	Flavanoids, Hue, OD280/OD315 Diluted Wine, Colour Intensity, Uronic Acid
`vscc-manly`-full	2	0.47	Flavanoids, Hue, OD280/OD315 Diluted Wine, Colour Intensity, Uronic Acid, Total Phenols
`skewvarsel` + MSN	3	0.78	Flavanoids, Hue, Proline, Colour Intensity, Alcohol, Uronic Acid, Malic Acid, Tartaric Acid
`skewvarsel` + Manly forward	3	0.73	Flavanoids, Hue, Proline, Colour Intensity, Alcohol, Uronic Acid, Malic Acid, Tartaric Acid
`skewvarsel` + Manly backward	2	0.46	Flavanoids, Hue, Proline, Colour Intensity, Alcohol, Uronic Acid, Malic Acid, Tartaric Acid

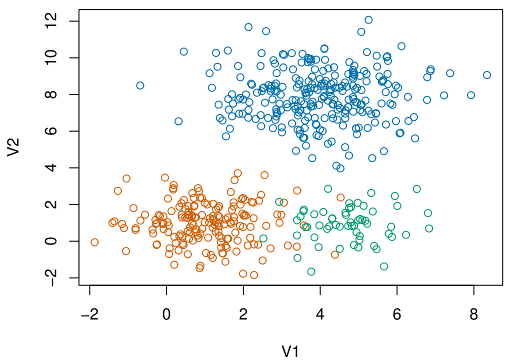


(a) True clustering on variables selected by `vscc-manly`-backwards.

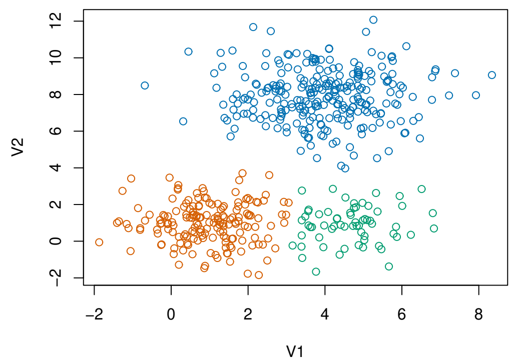


(b) Clustering by VSCC Manly B

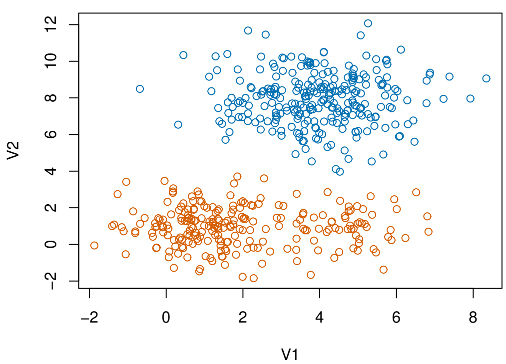
Figure 4.3: Variable selection and model fitting by `vscc-manly`-backwards on the Italian wine dataset.



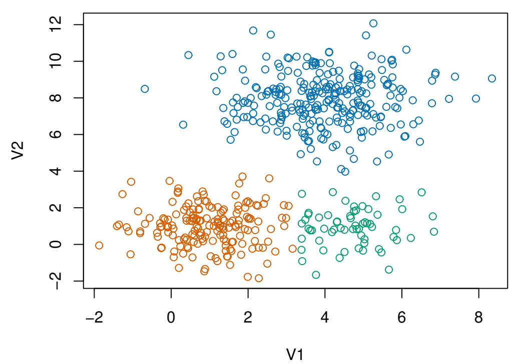
(a) True clusters (components) simulated data from three-component GMM.



(b) Clustering results for `vscc-manly`-forwards.



(c) Clustering results for `vscc-manly`-backwards



(d) Clustering results for `skewvarsel`+MSN.

Figure 4.4: Clustering results from simulated three-component GMM.

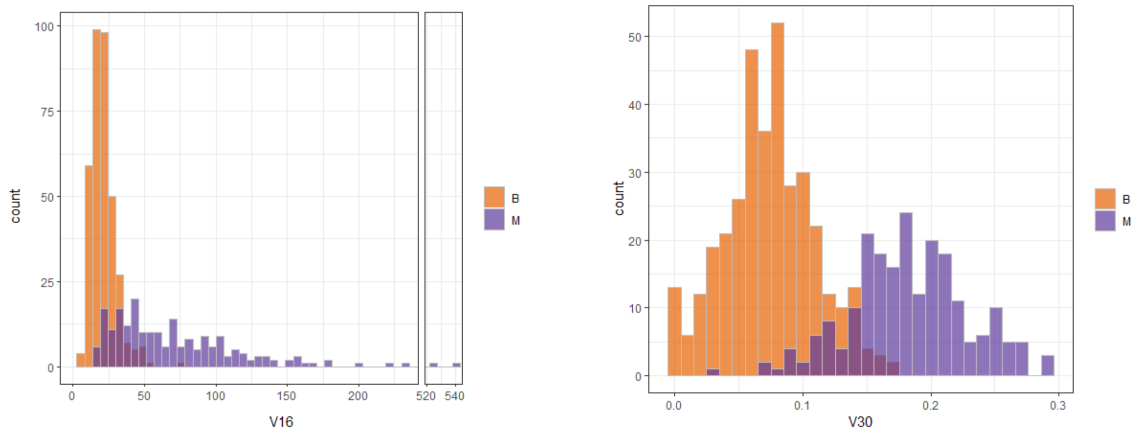
4.1.4 Breast Cancer Wisconsin (Diagnostic)

The breast cancer dataset comes from the UCI Machine Learning Repository (Dua and Graff, 2019). It contains 30 variables, two tumour types (benign and malignant), and 569 observations.

Table 4.4: Variable selection results for the breast cancer data.

Model	G	ARI	Variables
`vscc`	5	0.26	V8,V9,V10,V13,V18,V19,V20 V22,V26,V28,V29,V30
`clustvarsel`	4	0.39	V3,V5,V6,V8,V9,V13,V15,V16,V18 V19,V22,V23,V25,V26,V28,V29
`vscc-manly`-forward	2	0.50	V16
`vscc-manly`-backward	2	0.63	V30
`vscc-manly`-full	2	0.41	V5
`skewvarsel`+ MSN	5	0.33	V3, V6, V13, V16, V23, V26
`skewvarsel`+ Manly forward	4	0.45	V3, V6, V13, V16, V23, V26
`skewvarsel`+ Manly backward	3	0.36	V3, V6, V13, V16, V23, V26

From Table 4.4, we see that `vscc-manly` reduces the dimensions from 30 variables down to one, while selecting the correct number of clusters and obtaining the highest ARI of all methods tested. In particular, we see a large jump in performance, on all three measures, from `vscc` to its skewed counterpart. We do not see similar improvement in performance by the skewed extension of clustvarsel. Even upon fitting a Manly to the variables selected by `skewvarsel`, the performance in terms of G and ARI do not reach that of `vscc-manly` with backwards selection. This jump in performance from `vscc` to the `vscc-manly` is likely due to the strong skewness seen in some of the variables, as exhibited in Figure 4.5.



(a) Variables selected by `vscc-manly`-forwards.

(b) Variables selected by `vscc-manly`-backwards.

Figure 4.5: Breast cancer variables selected by `vscc-manly`.

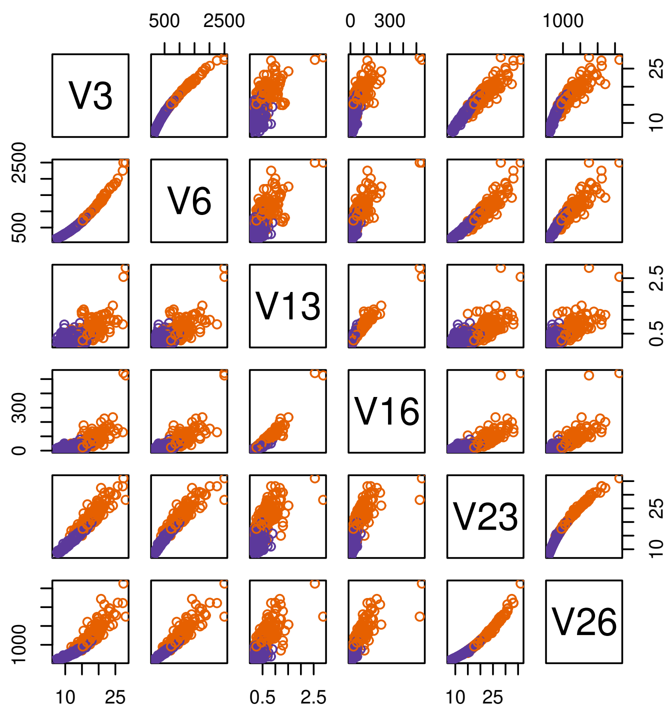


Figure 4.6: Breast cancer variables selected by `skewvarsel`.

4.2 Simulated Data Results

We simulated data from a three-component mixture of multivariate variance-gamma distributions 250 times. An example of this data can be found in Figure 4.7. To allow us to test for the effect of sample size on method performance, we ran our simulation at N $= 200, 500, \&1000$. Using a simulation is helpful as we can artificially create clustering and non-clustering variables to determine how well these methods select important variables and deselect unimportant ones. The simulation specifics are detailed in Table 4.5, where information on the clustering variables (V1 and V2), nonsense variables (V3 and V4) and the noisy variable (V5) can be found. To reduce the computational time, each model is fit to each simulated dataset only once.

Table 4.5: Simulated data information.

Simulation Data		
Clustering Variables		
$[X_{1g}, X_{2g}] \sim MVG(\mu_g, \Sigma_g, \alpha_g, \lambda_g, \psi_g)$		
$G = 1$	$G = 2$	$G = 3$
$\mu_1 = [2, 3]$	$\mu_2 = [5, 3]$	$\mu_3 = [5, 15]$
$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$
$\alpha_1 = [1, 4]$	$\alpha_2 = [4, 4]$	$\alpha_3 = [0.1, 0.1]$
$\lambda_1 = 4$	$\lambda_2 = 4$	$\lambda_3 = 3$
$\chi_1 = 0$	$\chi_2 = 0$	$\chi_3 = 0$
$\psi_1 = 8$	$\psi_2 = 8$	$\psi_3 = 6$
$p_1 = 0.4$	$p_2 = 0.4$	$p_3 = 0.2$
Nonsense Variables		
$X_3 \sim GIG(3, 0, 6)$		
$X_4 \sim GIG(1, 0, 2)$		
Noisy Variables		
$X_5 = 0.6 * V1 + 0.4 * Z$ where $Z \sim N(0, 5)$		

From Table 4.6, we see that the `vscc-manly`-backwards, `vscc-manly`-full, and `skewvarsel` algorithms perform the best in terms of G, ARI, and selecting the correct variables (V1 and V2) when $N = 500$ and $N = 1000$. For all three of these methods, performance improves as N increases. For both $N = 500$ and $N = 1000$, `skewvarsel` and `vscc-manly`-backwards select the correct variables every time. The performance of `skewvarsel` breaks down on all measures of performance when $N = 200$. We see a considerable standard deviation of ARI when $N = 200$ for `skewvarsel`, potentially suggesting instability at smaller sample sizes. Generally, these results agree with the real data results in that the skewed methods improve dimension reduction over their Gaussian counterparts when skewness is present.

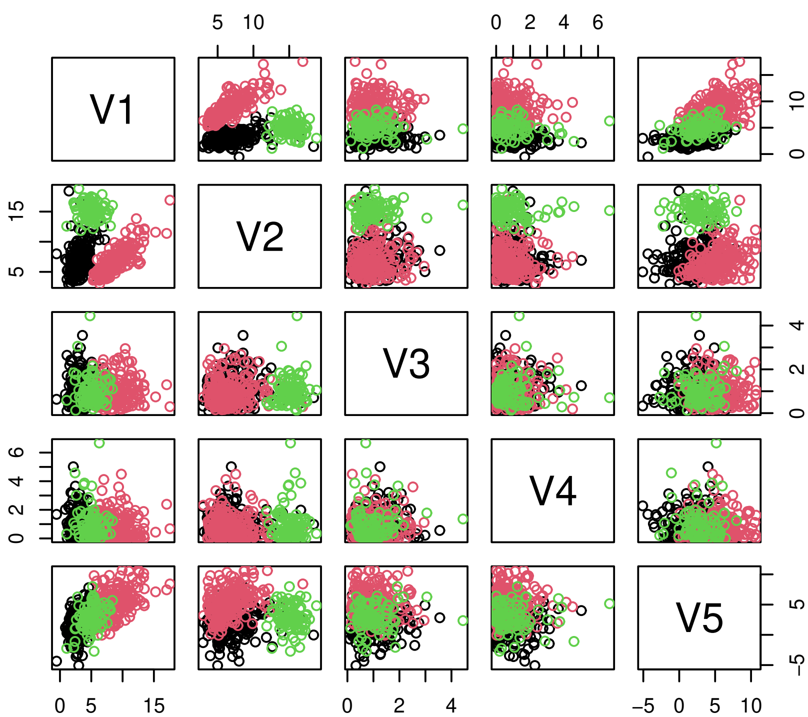


Figure 4.7: Example of simulated data when $N = 500$.

Table 4.6: Summary of simulation results.

Method	N	G	ARI	V1	V2	V3	V4	V5
`vscc`	200	3.9	0.82 (0.13)	250	250	5	34	230
	500	4.6	0.64 (0.1)	250	250	0	250	250
	1000	9	0.43 (0.02)	250	250	194	250	250
`clustvarsel`	200	4.7	0.62 (0.1)	250	250	0	227	0
	500	5.7	0.62 (0.05)	250	250	0	250	0
	1000	9	0.37 (0.007)	250	250	194	250	0
`vscc-manly`-forwards	200	3.3	0.89 (0.1)	246	247	6	4	10
	500	3.6	0.90 (0.13)	250	250	0	21	0
	1000	3.5	0.82 (0.15)	250	194	0	0	194
`vscc-manly`-backwards	200	3	0.95 (0.06)	248	250	2	4	5
	500	3	0.96 (0.01)	250	250	0	0	0
	1000	3	0.95 (0.000)	250	250	0	0	0
`vscc-manly`-full	200	3	0.94 (0.08)	247	249	5	1	4
	500	3	0.96 (0.01)	250	250	0	0	21
	1000	3	0.95 (0.002)	250	250	0	0	0
`skewvarsel`	200	2.6	0.59 (0.46)	155	156	0	95	3
+ MSN	500	3.14	0.92 (0.05)	250	250	0	0	0
	1000	3.4	0.92 (0.08)	250	250	0	0	0

Chapter 5

Discussion & Future Directions

In nearly all instances, we see the skewed extensions of common variable selection algorithms improving performance in the presence of skewness. This improvement in performance is seen in the selection of the number of clusters and ARI but more importantly, in the reduction of dimensions. In the AIS and breast cancer datasets, we see more effective dimension reduction by `vscc-manly` than `skewvarsel` in terms of the magnitude of dimension reduction and model fitting performance. For the banknote dataset, `vscc-manly` selects more variables than `skewvarsel` but also results in a better fitting model, regardless of the model fit to the `skewvarsel` results. The Italian wine dataset highlights the potential importance of utilizing methods designed for Gaussian clusters when appropriate.

There are instances where `vscc-manly` may select too many variables to account for some odd observations. For example, we see this in the AIS dataset when the backwards and full Manly extensions select variable Hg. This selection causes some boundary points between groups to switch clusters resulting in one less miss-classification.

Although adding this variable improves ARI, the goal of these algorithms is dimension reduction; as such, there may be instances in which smaller ARI is preferred if it is the result of a smaller selected set. One way to account for odd or hard-to-classify observations may be mixtures of contaminated transformation distributions. These component densities contain an inflated secondary component that allows for better modelling of outliers and heavy tails.

One downside to the skewed extensions is computational overhead. The `clustvarsel` and `skewvarsel` algorithms are naturally more computationally expensive due to their step-wise nature, with `skewvarsel` taking longer as more parameters need to be estimated. The `vscc` algorithm outperforms all methods on computational time as model fitting takes place only on the initial full set and the final sets of variables. This improvement in computational time extends into the skewed space when a full Manly is fit to the data. However, the algorithm slows down greatly under forward or backward transformation parameter selection due to the introduction of some inclusion/exclusion steps. This increase in computational time is heavily influenced by the structure of the clusters and the selection process used. For heavily skewed data, `vscc-manly` with backwards selection is much faster than its forwards counterpart. If only a few non-zero transformation parameters are necessary, `vscc-manly` with forwards selection would be much faster. Although more time-consuming than `vscc` or `vscc-manly`-full, `vscc-manly` with transformation parameter selection does tend to perform better than both in terms of ARI and dimension reduction. Thus, we suggest one performs some exploratory analysis on their data before selecting any of these methods to ensure that the algorithm selected is a good fit for their data and the computational overhead is justified. Additionally, the Manly transformation parameter selection is currently programmed in R (R Core Team, 2022) and could be sped up if programmed in a faster language. Computational time could be further reduced with parallelization of model fitting within each inclusion/exclusion step.

Bibliography

Adams, S. and Beling, P. A. (2019). A survey of feature selection methods for gaussian mixture models and hidden markov models. *Artificial Intelligence Review*, **52**(3), 1739–1779.

Andrews, J. L. and McNicholas, P. D. (2013). *vscc: Variable Selection for Clustering and Classification*. R package version 0.2.

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

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review*, **50**(2), 145–159.

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

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2), 211–243.

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, **43**(2), 176–198.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics. Chapman and Hall, London.

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

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**(6), 1149–1157.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York.

Gallaugher, M. P. B., McNicholas, P. D., Melnykov, V., and Zhu, X. (2020). Skewed distributions or transformations? modelling skewness for a cluster analysis.

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

Karlis, D. and Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, **19**(1), 73–83.

Lo, K. and Gottardo, R. (2012). Flexible mixture modeling via the multivariate t distribution with the box-cox transformation: an alternative to the skew-t distribution. *Statistics and computing*, **22**(1), 33–52.

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

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker Inc., New York.

McLachlan, G. J. and Peel, D. (2000). Mixtures of factor analyzers. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 599–606, San Francisco. Morgan Kaufmann.

McNicholas, P. D., ElSherbiny, A., McDaid, A. F., and Murphy, T. B. (2022). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.6.

McNicholas, S. M., McNicholas, P. D., and Browne, R. P. (2014). Mixtures of variance-gamma distributions. Arxiv preprint arXiv:1309.2695v2.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., Jager, P. L. D., and Mesirow, J. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, **106**, 8519–8524.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and 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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Scrucca, L. and Raftery, A. E. (2018). clustvarsel: A package implementing variable selection for gaussian model-based clustering in r. *Journal of Statistical Software*, **84**(1), 1–28.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 289–317.

Steinley, D. and Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, **73**, 125–144.

Steinley, D. and Brusco, M. J. (2011). K-means clustering and mixture model clustering: Reply to mclachlan (2011) and vermunt (2011). *Psychological Methods*, **16**(1), 89–92.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester.

Wallace, M. L., Buysse, D. J., Germain, A., Hall, M. H., and Iyengar, S. (2018). Variable selection for skewed model-based clustering: application to the identification of novel sleep phenotypes. *Journal of the American Statistical Association*, **113**(521), 95–110.

Zhu, X. and Melnykov, V. (2018a). Manly transformation in finite mixture modeling. *Computational Statistics & Data Analysis*, **121**, 190–208.

Zhu, X. and Melnykov, V. (2018b). *ManlyMix: An R Package for Model-Based Clustering with Manly Mixture Models*. R package version 0.1.14.