Variable Selection for Clustering via Manly Transformation

Elisa Du

Supervised by Dr.Paul McNicholas

McMaster University

April 9, 2020

Introduction

- In the era of big data, data sets are becoming more massive.
- High-dimensional data can lead to overparameterization when fitting models.
- This makes variable selection an important area of study.

Introduction (cont'd)

- Clustering is a form of unsupervised learning.
- Finite mixture models (McLachlan and Peel, 2000) have become a popular tool for clustering.
- Clusters in this context can be viewed as components within a mixture model (McNicholas, 2016).

Introduction (cont'd)

- Challenges in clustering with high-dimensional data include:
  - high computational intensity
  - reduced interpretability in the data set
  - high financial costs
  - at times unsatisfactory clustering performance of the model

  (Andrews and McNicholas, 2014).

- Variable selection techniques can improve clustering performance.
  - by removing noisy variables to improve the algorithm's ability to achieve more distinctive data group separation.

Introduction (cont'd)

- Much work has been done on model-based variable selection methods in *Gaussian* settings.
- But in many real-world data sets, variables deviate from normality.
  - Variable selection methods in Gaussian settings are no longer plausible.
- Here we propose and test an approach that enables *skewed* variable selection.

Finite Mixture Models

- Finite mixture models can be applied in three settings: model-based clustering, classification, and discriminant analysis (McNicholas, 2016).
- Here we focus on clustering.
- The density of a random vector $\mathbf{X}$, for all $\mathbf{x} \subset \mathbf{X}$, from a finite mixture distribution is written

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

where mixing proportions $\pi_g > 0$ and satisfy $\sum_{g=1}^{G} \pi_g = 1$, $f_1(\mathbf{x}|\boldsymbol{\theta}_g), ..., f_G(\mathbf{x}|\boldsymbol{\theta}_g)$ are the component densities that are usually of the same type, and $\boldsymbol{\nu} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G)$ is the vector of parameters with $\boldsymbol{\pi} = (\pi_1, ..., \pi_G)$.

Gaussian Finite Mixture Model

- A popular mixture model is the mixture of multivariate Gaussian distributions, with density written

$$f(\mathbf{x}|\boldsymbol{\nu}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2}$$

  where $\phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the multivariate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. (McNicholas, 2016).

- Next we introduce a family of models based on the Gaussian mixture model.

MCLUST Family of Mixture Models

- The total number of parameters of a p-dimensional random variable from a G-component Gaussian mixture model is equal to

$$G - 1 + Gp + \frac{Gp(p+1)}{2}. \tag{3}$$

- Banfield and Raftery (1993) proposed placing constraints on the elements of the eigen-decomposed components of the covariance matrix.

- The constraints gave rise to the family of eight Gaussian models known as the MCLUST family. (Fraley and Raftery, 2002).

Model Selection in MCLUST Family

- A well-established model selection criterion is the Bayesian information criterion (BIC; Schwarz, 1978).

- The MCLUST family of models (Fraley and Raftery, 2002) chooses the model with the highest BIC, here given

$$\text{BIC} = 2\ell(\hat{\boldsymbol{v}}) - \rho \log n, \tag{4}$$

where $\hat{\boldsymbol{v}}$ is the maximum likelihood estimate of $\boldsymbol{v}$, $\ell(\hat{\boldsymbol{v}})$ is the maximized log-likelihood, n is the number of observations, and ρ is the number of free parameters estimated in the model.

Membership Labels

- The true membership of observation $\mathbf{x_i}$ from component g is denoted as the indicator variable z_{ig}, i.e.

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{x_i} = (x_{i1}, ..., x_{ip}) \text{ belongs to cluster g,} \\ 0 & \text{otherwise.} \end{cases}$$

- "Soft" or "fuzzy" *a posteriori* predicted membership falling in the interval $[0, 1]$ is often used for interpretability, given

$$\hat{z}_{ig} := \frac{\hat{\pi}_g \phi\left(\mathbf{x_i} \middle| \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g\right)}{\sum_{h=1}^{G} \hat{\pi}_h \phi\left(\mathbf{x_i} \middle| \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h\right)} \tag{5}$$

for $i = 1, ..., n$ and $g = 1, ..., G$ (McNicholas, 2016).

Parameter Estimation

- The expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is commonly used to compute model parameter estimates.
- E-step updates $\hat{z}_{ig}$, and M-step updates parameter estimates.
  - The two steps are iterated until convergence, usually by some lack of progress criterion.
- We will introduce a variant of the EM algorithm later on...

Clustering Performance Measures

- The Rand index (ARI; Rand, 1971) measures the similarity between estimated and actual clustering membership.

$$RI = \frac{\text{number of pairwise agreements}}{\text{total number of pairs}}. \tag{6}$$

- The adjusted Rand Index(ARI; Hubert and Arabie, 1985) accounts for chance agreement between two random groupings.
  - Expected value of ARI = 0 indicates random clustering (Steinley, 2004).
  - ARI = 1 indicates perfect cluster agreement.

Variable Selection in Gaussian settings

- We introduce two variable selection techniques suitable in Gaussian settings (McNicholas, 2016).
- *clustvarsel* selects variables by comparing models using approximate Bayes factors (Kass and Raftery, 1995).
- Variable selection for clustering and classification (VSCC) technique has the aim of "simultaneously minimizing the 'within-group' variance and maximizing the 'between-group' variance" (Andrews and McNicholas, 2014).

Skewed Variable Selection

- Skewed variables are encountered in research in sleep (Wallace et al., 2017), drug-screening (Lo, Brinkman, and Gottardo, 2008), and facial recognition (Wang et al., 2019).
- Fitting Gaussian mixture models to skewed data can result in poor clustering recovery or overfitting.
- Two approaches to model skewness are:
  1. choosing components that use asymmetric distributions to improve fit, and
  2. transforming the data to near-normality.

Approach 1: *Skewvarsel*

- Wallace et al. (2017) recently introduced the *skewvarsel* technique for skewed variable selection.
- Recall *clustvarsel* (Raftery and Dean, 2006; Scrucca and Raftery, 2018) is applicable in Gaussian settings.
- *skewvarsel* is a step-wise regression algorithm that extends *clustvarsel* to the multivariate skew normal distribution. (Pyne et al., 2009; Azzalini and Valle, 1996)

Approach 2: Manly Transformation

- A widely known transformation is the Box-Cox power transformation (Box and Cox, 1964).
  - It has the drawback of only handling positive data values.
- Manly exponential transformation (Manly, 1976) can handle data ranging from $-\infty$ to $+\infty$.
- In the univariate case, for the scalar variable x, Manly transformation is given by

$$\mathrm{y} = \begin{cases} \frac{\mathrm{e}^{\lambda \mathrm{x}} - 1}{\lambda} & \lambda \neq 0, \\ \mathrm{x} & \lambda = 0, \end{cases}$$

where λ is the transformation parameter and y is the transformed variable.

Multivariate Manly Transformation

- Zhu and Melnykov (2018) extend the Manly transformation to the multivariate scenario, where there are n p-dimensional data vectors $\mathbf{x_1}, ..., \mathbf{x_n}$. Assume there exists a transformation vector $\boldsymbol{\lambda}_g = (\lambda_{g1}, ..., \lambda_{gp})'$ for component $g = 1, ..., G$, so that the transformed vector is written

$$\mathbf{Y}_g = \left(\frac{\mathbf{e}^{\lambda_{g1}\mathbf{x}_1} - 1}{\lambda_{g1}}, ..., \frac{\mathbf{e}^{\lambda_{gp}\mathbf{x}_p} - 1}{\lambda_{gp}} \right) \sim N_p \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right). \quad (7)$$

- This led to the **Manly mixture model** used to model skew data...

The Manly Mixture model

- Each component of the Manly mixture model (Zhu and Melnykov, 2018) can be obtained via a back-transformation from the Gaussian distribution to the original skewed data.

- The density of the Manly mixture model is written

$$g\left(\mathbf{x}\middle|\boldsymbol{\nu}\right)=\sum_{g=1}^{G}\pi_g\phi\left(\mathcal{M}\left(\mathbf{x},\boldsymbol{\lambda}_g\right)\middle|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)\exp\{\boldsymbol{\lambda}_g^{'}\boldsymbol{x}\},\quad(8)$$

where $\mathcal{M}\left(\mathbf{x},\boldsymbol{\lambda}_g\right)\equiv\mathbf{Y}_g$ is the original data $\boldsymbol{\mathcal{X}}=\mathbf{x_1},...,\mathbf{x_n}$ transformed to normality, $\phi\left(\cdot\middle|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)$ is the multivariate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\lambda}_g$ is the Manly transformation parameter of $\boldsymbol{\mathcal{X}}$ for the gth component.

Manly-EM Introduction

- Zhu and Melnykov (2018) proposed a variant of the EM algorithm to compute parameter estimates of the Manly mixture model.
- Here we refer to this algorithm as *Manly-EM*.
- The likelihood of Manly mixture model can be written by

$$\mathcal{L}(\boldsymbol{\nu}) = \prod_{i=1}^{n} \left(\sum_{g=1}^{G} \pi_g \phi\Big(\mathcal{M}(\mathbf{x}, \boldsymbol{\lambda}_g) \Big| \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \Big) \exp\{\boldsymbol{\lambda}_g' \boldsymbol{x}\} \right), \tag{9}$$

where $\mathcal{M}(\mathbf{x}, \boldsymbol{\lambda}_g)$ is the transformed data, $\phi\Big(\cdot \Big| \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \Big)$ is the multivariate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\lambda}_g$ is the Manly transformation parameter of $\boldsymbol{\mathcal{X}}$ for the gth component.

Manly-EM algorithm

- In the E-step, the conditional expectation of the complete-data log-likelihood is computed, also known as the $\mathcal{Q}$-function, given by

$$\mathcal{Q}\left(\boldsymbol{\nu}\middle|\hat{\boldsymbol{\nu}},\mathbf{x}\right)=\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\left[\log\left\{\pi_g\phi\left(\mathcal{M}\left(\mathbf{x}_i,\boldsymbol{\lambda}_g\right)\middle|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)\right\}+\boldsymbol{\lambda}_g^{'}\mathbf{x}_i\right], \tag{10}$$

which simplifies to updating $\hat{z}_{ig}$, the *a posteriori* membership probability for each component, given by

$$\hat{z}_{ig}=\frac{\hat{\pi}_g\phi\left(\mathcal{M}\left(\mathbf{x}_i,\hat{\boldsymbol{\lambda}}_g\right)\middle|\hat{\boldsymbol{\mu}}_g,\hat{\boldsymbol{\Sigma}}_g\right)\exp\{\hat{\boldsymbol{\lambda}}_g^{'}\mathbf{x}_i\}}{\sum_{h=1}^{G}\hat{\pi}_h\phi\left(\mathcal{M}\left(\mathbf{x}_i,\hat{\boldsymbol{\lambda}}_h\right)\middle|\hat{\boldsymbol{\mu}}_h,\hat{\boldsymbol{\Sigma}}_h\right)\exp\{\hat{\boldsymbol{\lambda}}_h^{'}\mathbf{x}_i\}}. \tag{11}$$

Manly-EM algorithm (cont'd)

- The M-step computes the parameter estimates by maximizing the $\mathcal{Q}$-function with respect to each of π_g, $\boldsymbol{\mu}_g$, and $\boldsymbol{\Sigma}_g$, giving

$$\hat{\pi}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{n}, \qquad (12) \qquad \hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} \mathcal{M}\left(\mathbf{x}_i, \hat{\lambda}_g\right)}{\sum_{i=1}^{n} \hat{z}_{ig}}, \qquad (13)$$

$$\hat{\boldsymbol{\Sigma}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} \left(\mathcal{M}\left(\mathbf{x}_i, \hat{\lambda}_g\right) - \hat{\boldsymbol{\mu}}_g\right)\left(\mathcal{M}\left(\mathbf{x}_i, \hat{\lambda}_g\right) - \hat{\boldsymbol{\mu}}_g\right)'}{\sum_{i=1}^{n} \hat{z}_{ig}}. \qquad (14)$$

- Since the closed-form solution for $\hat{\boldsymbol{\lambda}}_{\boldsymbol{g}}$ is not available, Nelder-Mead numerical optimization (Nelder and Mead, 1965) is used.

Manly-EM algorithm (cont'd)

- Nelder-Mead method (Nelder and Mead, 1965) find the values that optimize a multidimensional unconstrained function without requiring any derivative information.
- To compute $\hat{\boldsymbol{\lambda}}_{\boldsymbol{g}}$, the component-wise $\mathcal{Q}$-function is maximized using Nelder-Mead optimization, written

$$\mathcal{Q}_g\left(\boldsymbol{\lambda_g}\middle|\hat{\boldsymbol{\nu}}\right)(\boldsymbol{\lambda}_g) = \sum_{i=1}^{n} \hat{z}_{ig}\left\{\log\phi\left(\mathcal{M}\left(\mathbf{x}_i, \boldsymbol{\lambda}_g\right)\middle|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) + \boldsymbol{\lambda}_g' \mathbf{x}_i\right\} + const. \tag{15}$$

- Note the relation between $\mathcal{Q}$ and $\mathcal{Q}_g$:

$$\mathcal{Q}\left(\boldsymbol{\nu}\middle|\hat{\boldsymbol{\nu}}, \mathbf{x}\right) = \sum_{g=1}^{G}\left[\sum_{i=1}^{n} \hat{z}_{ig}\log\hat{\pi}_g + \mathcal{Q}_g\right]. \tag{16}$$

Manly-EM: initialization and convergence

- Convergence criterion is the relative difference of $\mathcal{Q}$-function values between consecutive iterations, i.e.

$$\mathcal{Q}^{(k)} - \mathcal{Q}^{(k-1)} < \epsilon, \tag{17}$$

  where $\mathcal{Q}^{(k)}$ is the conditional expectation of complete-data log-likelihood at iteration k.

- Initialization of $\hat{z}_{ig}$ is via K-means (Macqueen, 1967) clustering, and $\hat{\boldsymbol{\lambda}}_g$ is initialized with value of 0.1 for each of the p variables.

VSCC algorithm

- The central idea of VSCC is to find the variables that minimize the within-group variance and maximize the between-group variance (Andrews and McNicholas, 2014).

- First, the within-group variance for each variable $j = 1, ..., p$ is calculated, written

$$W_j = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \left(\mathrm{x}_{ij} - \mu_{gj}\right)^2}{n}, \tag{18}$$

  where x_{ij} is observation i on variable j, μ_{gj} is the mean of variable j in group g, n is the number of observations, and z_{ig} is the group membership indicator variable.

- The data is assumed to have been standarized to have mean 0 and variance 1.

VSCC algorithm (cont'd)

- The first variable selected is one with minimum W_j.
- To select remaining variables, the within-group variance and between-variable correlation is used as criterion:

$$|\rho_{jr}| < (1 - W_j)^m , \tag{19}$$

where ρ_{jr} is the correlation between variables, for all $r \in V$ where V is the space of currently selected variables, and $m \in \{1, ..., 5\}$ is fixed.

- Up to 5 distinct variable subsets can be selected.

VSCC algorithm(cont'd)

- *Mclust* is used to perform clustering for the (up to 5) distinct variable subsets.

- The best variable subset is taken to be the one that minimizes the total model uncertainty in the fuzzy clustering matrix $\hat{z}_{ig}$, which is

$$n - \sum_{i=1}^{n} \max\{\hat{z}_{ig}\}.. \tag{20}$$

VSCC limitation

- VSCC can only be applied for Gaussian scenarios, so it falls short in identifying skewed clusters in data sets that deviate from normality.
- It is this limitation that motivates the variable selection approach proposed in the next section.

Manly-VSCC

- We will refer to the proposed skewed-variable selection approach as *Manly-VSCC*.
- Pseudocode below:
  1. *Manly-EM* is used to find parameter estimates for the Manly mixture model, and for our purpose, we use it to obtain $\hat{z}_{ig}$ and $\hat{\boldsymbol{\lambda}}_g$.
  2. Based on $\hat{z}_{ig}$, multivariate Manly transformation is applied to the observation $\mathbf{x}_i$ identified to belong to component g. This results a p-variate normally distributed data set.
  3. *VSCC* is applied to the transformed data to obtain the best variable subset.
  4. Clustering results with *mclust* is reported for the best subset.

Applications

- **Manly-VSCC** was evaluated by comparing its clustering results with those from three other approaches:
  - **Mclust** (no variable selection)
  - **VSCC**
  - **skewvarsel**.
- *Mclust* was used as the model-based clustering algorithm for techniques involving *VSCC*.
- *skewvarsel* used the mixture of generalized hyperbolic distributions (McNeil, Frey, and Embrechts, 2005) to obtain clustering results.

AIS data set

- The Australian Institute of Sport (AIS) data set (Cook and Weisberg, 1994) contains information on 202 athletes, 102 male and 100 female.
- 11 (numeric) out of the 13 variables were used to construct clustering models.
- The goal of the analysis is to cluster the athletes into male and female groups.

AIS data set (cont'd)

- *Manly-EM* clustering results has 20 misclassifications.

Table: True group memberships (Male, Female) against predicted group memberships (1,2) using Manly-EM. ARI = 0.64.

	1	2
M	89	13
F	7	93

AIS data set (cont'd)

- The *skewvarsel* solution picked the correct number of components (G = 2), with 8 misclassifications.

	1	2	3
M	7	8	87
F	90	3	7

(a) Manly-VSCC

	1	2	3	4
M	1	2	88	11
F	56	42	0	2

(b) VSCC

	1	2	3	4
M	0	38	59	5
F	55	12	0	33

(c) Mclust

	1	2
M	94	8
F	0	100

(d) skewvarsel

Table: True group memberships (Male, Female) against predicted group memberships (1,2) for the AIS data set.

AIS data set (cont'd)

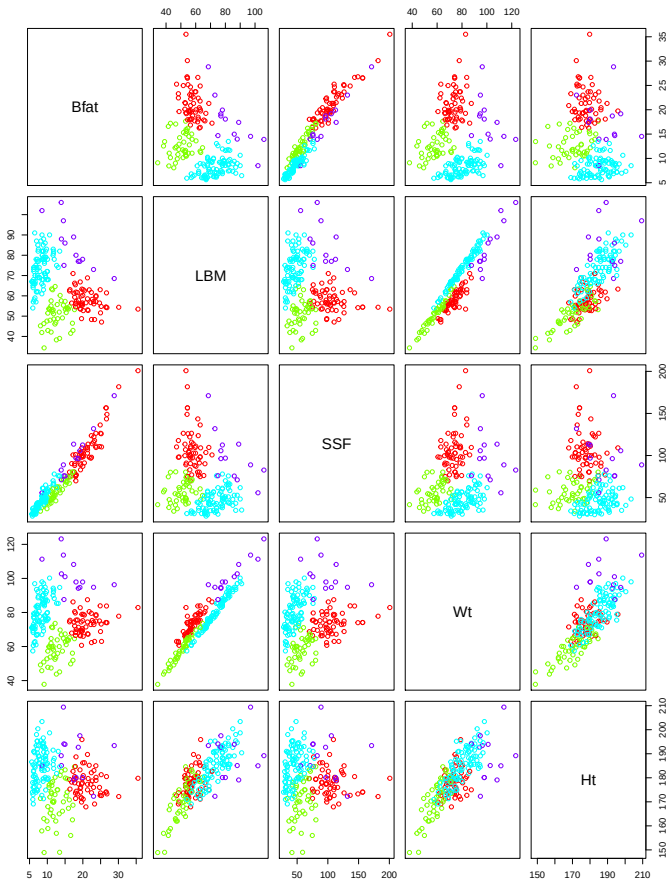

Figure: VSCC solution.

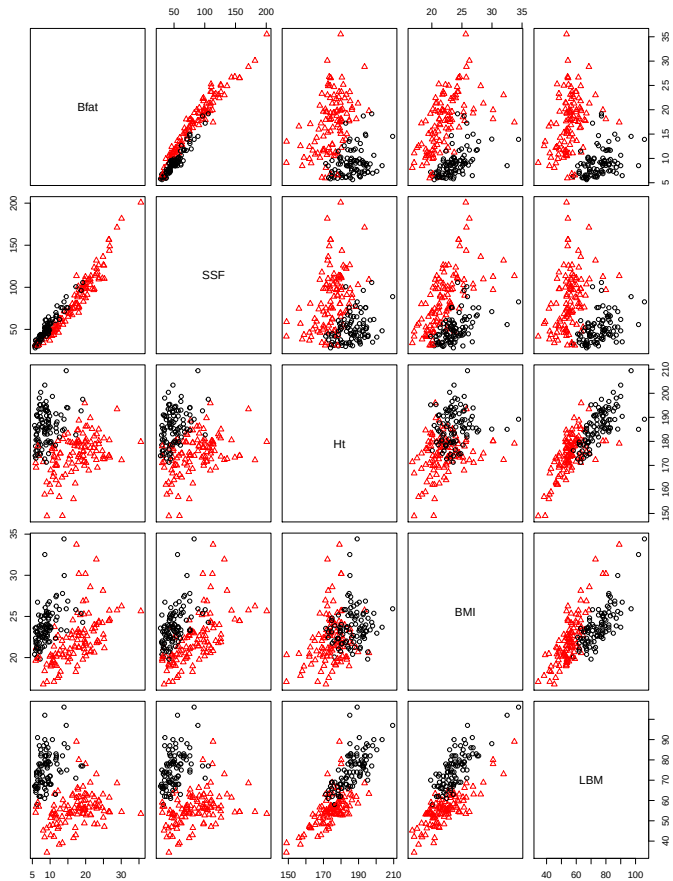

Figure: skewvarsel.

AIS data set (cont'd)

Table: Summary model and clustering results using Manly-VSCC,VSCC, Mclust, and skewvarsel methods for AIS data set.

	G	ARI
Manly-VSCC	3	0.65
VSCC	4	0.61
Mclust	4	0.39
skewvarsel	2	0.85

- *skewvarsel* results in the highest ARI of 0.85, followed by the second-highest ARI of the proposed *Manly-VSCC* approach, which is 0.65.
- This data set is an example of how variable selection can improve clustering performance.

Swiss bank notes data

- The Swiss bank notes data set (Fraley, Raftery, and Scrucca, 2016) includes 6 measurements made on 200 bank notes, including 100 genuine and 100 counterfeit.
- The clustering results using the *Manly-EM* algorithm is given. The algorithm demonstrated good clustering performance, with ARI of 0.85 and 8 misclassifications total.

Table: True group memberships (counterfeit, genuine) against predicted group memberships (1,2) using Manly-EM. ARI = 0.85.

	1	2
counterfeit	97	3
genuine	5	95

Swiss bank notes data (cont'd)

- *skewvarsel* performs the best, choosing the correct number of groups with only 1 misclassification.
- However, *Manly-VSCC* performs clustering poorly.

	1	2	3	4	5	6
counterfeit	0	2	1	32	31	34
genuine	33	33	29	4	1	0

(a) Manly-VSCC

	1	2	3
counterfeit	15	0	85
genuine	1	99	0

(b) VSCC

	1	2	3
counterfeit	16	0	84
genuine	2	98	0

(c) Mclust

	1	2
counterfeit	100	0
genuine	1	99

(d) skewvarsel

Swiss bank notes data (cont'd)

- *Skewvarsel* has the best clustering performance (ARI = 0.98).
- *VSCC* and *Mclust* have good clustering recovery.
- The unusually low ARI of *Manly-VSCC* indicates it falters for this data set.

Table: Summary model and clustering results for Swiss bank notes data.

	G	ARI
Manly-VSCC	6	0.28
VSCC	3	0.86
Mclust	3	0.84
skewvarsel	2	0.98

Italian Olive Oils data set

- The Italian Olive Oils data set (Forina et al., 1983; Forina and Tiscornia, 1982) contains 8 fatty acid measurements on 572 Italian olive oil samples originating from 3 regions(323 samples from Southern Italy, 98 from Sardinia, and 151 from Northern Italy).
- Poor clustering results is demonstrated by *Manly-EM*, with low ARI of 0.41.

Table: True group memberships (Southern Italy, Sardinia, Northern Italy) against predicted group memberships (1,2,3) using Manly-EM.

	1	2	3
S.Italy	0	121	202
Sardinia	97	0	1
N.Italy	88	63	0

Italian Olive Oils data set (cont'd)

- Scatterplot matrices of *VSCC* and *skewvarsel* are shown.

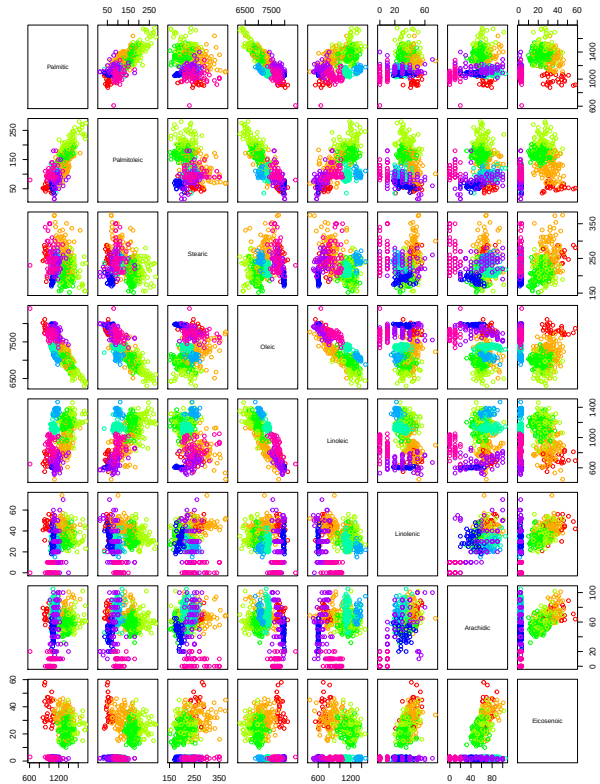

Figure: VSCC solution.

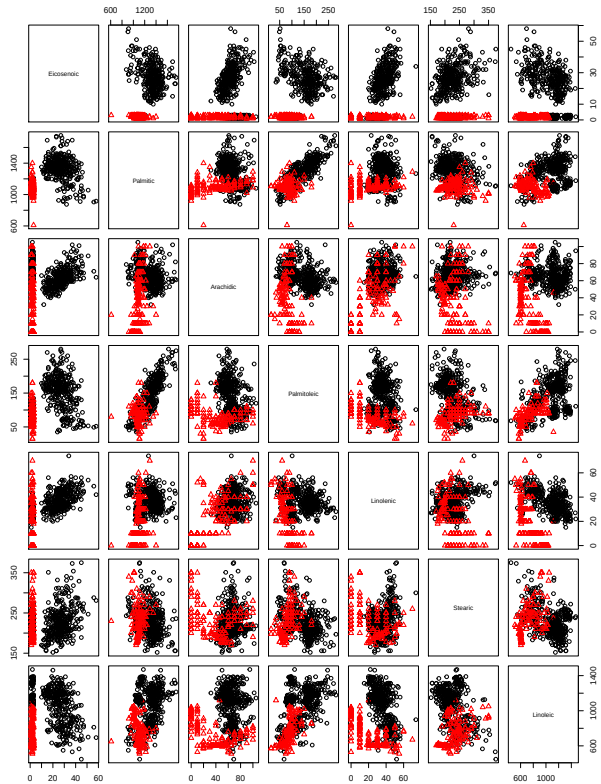

Figure: skewvarsel solution.

Italian Olive Oils data set(cont'd)

- *skewvarsel* outperform the 3 other approaches in having the highest ARI (0.63) using a two-cluster solution.
- *Manly-VSCC* and *VSCC* perform roughly on par, but neither perform as well as *Mclust*.

Table: Summary model and clustering results on Italian Olive Oils data set.

	G	ARI
Manly-VSCC	6	0.41
VSCC	9	0.35
Mclust	6	0.56
skewvarsel	2	0.63

Summary

- We proposed and evaluated a skewed-variable selection approach under a clustering framework.
- The approach is tested with three real data sets, of which one performed well but two others did poorly.
- The poor performance can be attributed to one or more of the following:
  - Initialization method causing unstable *Manly-EM* performance.
  - Inaccurate parameter estimated by the *Manly-EM* algorithm, which fails to transform data to near-normality.
  - Issues in using Nelder-Mead optimization to update the skewness transformation parameters.

Future approach to investigate

- *skewvarsel* demonstrates promising performance across the three data sets.
- Recall that *skewvarsel* selects for skewed variables using the multivariate skew normal distribution (Wallace et al., 2017).
- We propose a similar approach using the variance-gamma mixture model.
- The mixture of variance-gamma (VG) distributions arises from placing parameter restrictions on the generalized hyperbolic distribution (McNeil, Frey, and Embrechts, 2005).

Using variance-gamma distance for skewed-VSCC?

- Two candidate distances identified from the variance-gamma density are:

$$\sqrt{\left(\psi + \boldsymbol{\alpha}' \boldsymbol{\Sigma}^{-1} \alpha\right) \delta\left(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}\right)} \tag{21}$$

and

$$\left\{ \left(\boldsymbol{\mu} - \mathbf{x}\right)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} \right\} \tag{22}$$

where (21) is a modified Mahalanobis distance that accounts for both skewness and concentration, and (22) is one that accounts for skewness only.

- As future direction, we can derive within-group variance expressions for each of (21) and (22) to account for skewness and arrive at a variable selection approach for asymmetric distributions.

Acknowledgements

I would like to thank Dr.McNicholas for his encouragement and guidance throughout the course of my thesis.

References I

Andrews, J. L. and P. D. McNicholas (2014). "Variable Selection for clustering and classification". In: *Journal of Classification* 31.2, pp. 136–153.

Azzalini, A. and A. D. Valle (1996). "Multivariate skew-normal distribution". In: *Biometrika* 83, pp. 715–726.

Banfield, J. D. and A. E. Raftery (1993). "Model-based Gaussian and non-Gaussian clustering". In: *Biometrics* 49, pp. 803–821.

Box, G. E. and D. R. Cox (1964). "An analysis of transformations". In: *Journal of the Royal Statistical Society* 26, pp. 211–252.

Cook, D. and S. Weisberg (1994). *An Introduction to Regression Graphics*. New York: John Wiley & Sons.

References II

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B* 39.1, pp. 1–38.

Forina, M. and E. Tiscornia (1982). "Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content". In: *Annali di Chimica* 72, pp. 143–155.

Forina, M. et al. (1983). *Classification of olive oils from their fatty acid composition*. London: Applied Science Publishers.

Fraley, A., E. Raftery, and L. Scrucca (2016). *mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*. R package version 5.2.

References III

Fraley, C. and A. E. Raftery (2002). "Model-based clustering,discriminant analysis, and density estimation". In: *Journal of the American Statistical Association* 97, pp. 611–631.

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

Kass, R. E. and A. E. Raftery (1995). "Bayes factors". In: *Journal of the American Statistical Association* 90, pp. 773–795.

Lo, K., R. R. Brinkman, and R. Gottardo (2008). "Automated gating of flow cytometry data via robust model-based clustering". In: *Cytometry* 73, pp. 321–332.

References IV

Macqueen, J. (1967). "Some methods for classification and analysis of multivariate observations". In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Manly, B. F. J (1976). "Exponential data transformations". In: *Journal of the Royal Statistical Society* 25, pp. 37–42.

McLachlan, G. J. and D. Peel (2000). *Finite mixture models*. John Wiley & Sons.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton: Princeton University Press.

McNicholas, P. D. (2016). *Mixture Model-Based Classification*. CRC Press.

References V

Nelder, J. A. and R. Mead (1965). "A simplex method for function minimization". In: *The Computer Journal* 7, pp. 308–313.

Pyne, S. et al. (2009). "Proceedings of the National Academy of Sciences in the United States of America". In: vol. 106. National Academy of Sciences.

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

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical Association* 66, pp. 846–850.

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

References VI

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

Steinley, D. (2004). "Properties of the Hubert-Arabie adjusted Rand index". In: *Psychological Methods* 9, pp. 386–396.

Wallace, M. L. et al. (2017). "Variable selection for skewed model-based clustering: Application to the identification of novel sleep phenotypes". In: *Journal of the American Statistical Association* 113, pp. 95–110.

Wang, P. et al. (2019). "Deep class-skewed learning for face recognition". In: *Neurocomputing* 363, pp. 35–45.

References VII

Zhu, X. W. and V. Melnykov (2018). "Manly transformation in finite mixture modeling". In: *Computational Statistics and Data Analysis* 121, pp. 190–208.

Parameter Estimation for Gaussian Mixture

- For the mixture of multivariate Gaussians, the clustering log-likelihood is given by

$$\log \mathcal{L}(\boldsymbol{v}) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g \phi\left(\mathbf{x_i} \middle| \hat{\mu}_{\mathbf{g}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}\right). \tag{23}$$

Manly-EM pseudocode

initialize $\hat{z}_{ig}$ and $\hat{\boldsymbol{\lambda}}_g$
initialize $\hat{\pi}_g$, $\mathcal{M}(\mathbf{x}, \hat{\boldsymbol{\lambda}}_g)$, $\hat{\boldsymbol{\mu}}_g$, $\hat{\boldsymbol{\Sigma}}_g$
while not converged
update $\hat{z}_{ig}$
update $\hat{\boldsymbol{\lambda}}_g$ via Nelder-Mead optimization of $-\mathcal{Q}_g$
update $\mathcal{M}(\mathbf{x}, \hat{\boldsymbol{\lambda}}_g)$
update $\hat{\pi}_g$, $\hat{\boldsymbol{\mu}}_g$, $\hat{\boldsymbol{\Sigma}}_g$
check convergence criterion
end while

AIS data set - results

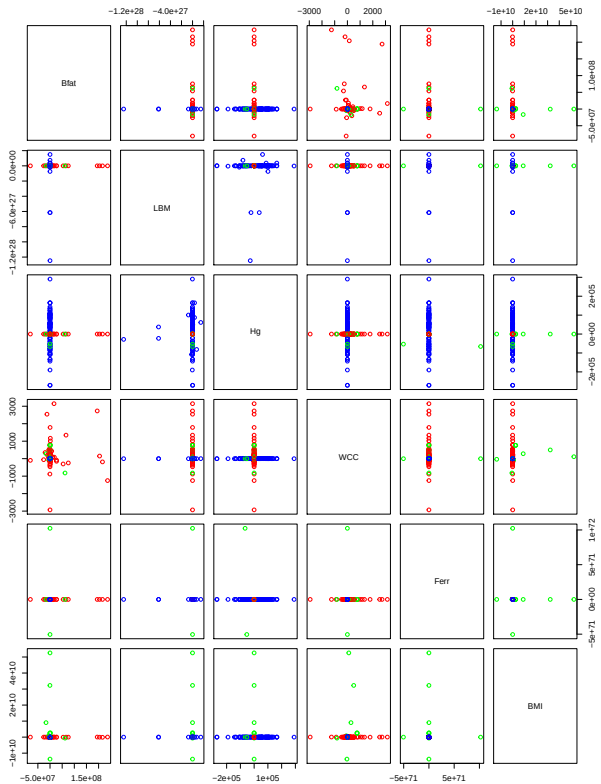

Figure: scatterplot matrix of the VSCC 5 selected variables on the AIS data, where the colours correspond to the different groups found via the method.

AIS data set (cont'd)

- *Manly-VSCC* retained 1 more variable than the other approaches.
- *Both VSCC* techniques chose the quadratic relation as the variable subset that minimizes clustering model uncertainty.

Table: Summary results on variable selection for AIS data set.

	num. var. selected	names of var.selected
Manly-VSCC (Quadratic)	6	Bfat,Hg,WCC,Ferr,BMI,LBM
VSCC(Quadratic)	5	Bfat,SSF,Ht,BMI,LBM
skewvarsel	5	Bfat,SSF,Ht,Wt,LBM

Swiss bank note data set - results

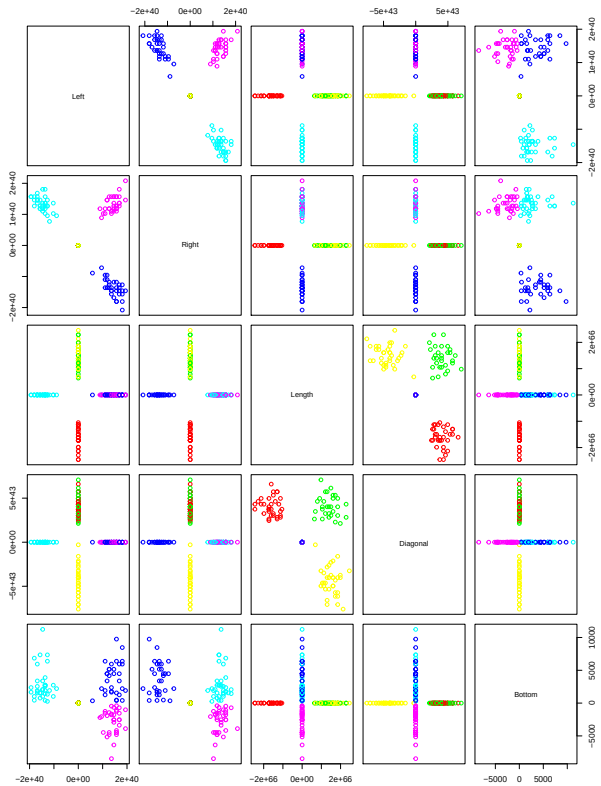

Figure: scatterplot matrix of the 5 selected variables using Manly-VSCC on the Swiss bank notes data.

Swiss bank notes data

- *Manly-VSCC* retains one more variable than the other two approaches using the linear relation.
- *VSCC* chooses the cubic relation, and chose the same number of variables as *skewvarsel*.

Table: Summary results on variable selection for Swiss bank notes data set.

	num. var. selected	names of var.selected
Manly-VSCC(Linear)	5	Length, Left, Right, Bottom, D
VSCC(Cubic)	4	Top,Right,Bottom,Diagon
skewvarsel	4	Top,Left,Bottom,Diagona

Future Direction for *Manly-EM*

- There are currently no constraints placed on the covariance matrix components of the Manly mixture model.
- A future direction would be to include in *Manly-EM* the 14 parameterizations of the within-group covariance matrix $\boldsymbol{\Sigma}_g$, as in for the *Mclust* family (Fraley, Raftery, and Scrucca, 2016).