2018

Variable Selection Techniques for Clustering on the Unit Hypersphere

Damon Bayer
South Dakota State University

VARIABLE SELECTION TECHNIQUES FOR CLUSTERING ON THE UNIT HYPERSPHERE

BY

DAMON BAYER

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Mathematics

Specialization in Statistics

South Dakota State University

2018

VARIABLE SELECTION TECHNIQUES FOR CLUSTERING ON THE UNIT HYPERSPHERE

DAMON BAYER

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Mathematics with a Specialization in Statistics degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Semhar Michael, Ph.D.
Thesis Advisor Date

Kurt Cogswell, Ph.D.
Head, Department of Mathematics & Statistics Date

Dean, Graduate School Date

ACKNOWLEDGMENTS

This work would not have been possible without the the support of Dr. Semhar Michael, whose mentorship and guidance have helped me thrive in my graduate studies. I am also grateful to her for the financial support I have received throughout my studies as an Arnold K. Skeie E-Commerce Analytics Graduate Fellow and as a Daschle Student Fellow. I would also like to thank Dr. Christopheer Saunders for his work as a part of my graduate committee and the many illuminating conversations shared inside and outside the classroom. Additionally, I extend my gratitude to Dr. Kurt Cogswell for his guidance during my undergraduate career, encouraging me to pursue graduate education, and offering me financial assistance as a Graduate Teaching Assistant to support my studies. Finally, I wish to acknowledge Dr. Jessice Meendering for serving as the Graduate Faculty Representative on my committee.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ABSTRACT

VARIABLE SELECTION TECHNIQUES FOR CLUSTERING ON THE UNIT HYPERSPHERE

DAMON BAYER

2018

Mixtures of von Mises-Fisher distributions have been shown to be an effective model for clustering data on a unit hypersphere, but variable selection for these models remains an important and challenging problem. In this paper, we derive two variants of the expectation-maximization framework, which are each used to identify a specific type of irrelevant variables for these models. The first type are noise variables, which are not useful for separating any pairs of clusters. The second type are redundant variables, which may be useful for separating pairs of clusters, but do not enable any additional separation beyond the separability provided by some other variables. Removing these irrelevant variables is shown to improve cluster quality in simulated as well as benchmark datasets.

1 INTRODUCTION

Model-based clustering uses mixtures of distributions to identify subpopulations in a dataset. Several distributions have been proposed for the model-based clustering of data on a hypersphere. Among these are mixtures of von Mises-Fisher distributions [1], mixtures of Watson distributions [3], mixtures of inverse stereographic projections of multivariate normal distributions [7], and mixtures of Poisson kernel distributions [8]. Text data is a common example of data modeled on the hypersphere, with normalized vectors of word counts being used to represent documents. Mixtures of multinomial and Bernoulli distributions have been used for clustering these documents [15, 24]. A comparison of these two to the previously mentioned mixtures of von Mises-Fisher distributions is presented in [29]. Various non-generative methods have also been proposed for clustering spherical data, including several modifications of the popular k-means clustering method [13, 28].

It is well known that the addition or presence of irrelevant variables in a dataset degrades the clustering performance [16, 17]. Additionally, performing dimension reduction as part of the preprocessing of dataset can degrade clustering quality [4]. For this reason, the study of variable selection in the context of clustering is important. Recently, several authors have studied this topic, primarily in the context of finite mixtures of multivariate normal distributions. The study of variable selection for model-based clustering has primarily followed two perspectives.

The first are methods which make explicit assumptions about the relationships between relevant clustering variables, irrelevant clustering variables, and mixture membership. The complexity of these assumed relationships has evolved over time, with Law et al. [11] first assuming independence between the relevant and irrelevant variables. Raftery and Dean propose an approach analogous to stepwise regression and assume irrelevant variables are condition-

ally independent of the cluster labels given the relevant variables [21]. In this approach, the irrelevant variables are assumed to be explained by the relevant variables. Maugis et al. [14] expand on this concept by further separating the set of irrelevant variables into those which are redundant and can be explained by the relevant variables, and those which are noise and cannot be explained by the relevant variables.

Other variable section methods for model-based clustering use regularization methods, wherein some penalty term is added to the likelihood function used for maximization. Penalized model-based clustering was first used in [19] to model mixtures of multivariate normal distributions with common isotropic covariance matrices. An L_1 penalty is used to obtain a sparse solution with many mean parameters being pushed to zero. This work has further been expanded in [25], which introduced an L_∞ penalty term to reduce the maximum mean for each cluster in order to shrink more mean parameters to zero. A pairwise fusion penalty was used in [9] to identify which variables can discriminate which pairs of clusters. The restriction of a common isotropic covariance matrix is relaxed in [27] and abandoned fully in [30] to develop penalized clustering with no constraints on covariance matrices. Recently, a hybrid approach was proposed in [6], which first uses the method proposed in [30] to rank the variables. The ranked variables are then processed in order for being noise or redundant by the method discussed in [14]. This hybrid method overcomes some of the computational challenges faced by the variable selection algorithms that follow from [11].

Variable selection methodologies also exist specifically for text data. The most common among these are the removal of "stop words," common words in a language (a, the, is) which are presumed to not provide relevant information for some analysis tasks, and variable selection by term frequency-inverse document (tf-idf) frequency, a measure of a word's concentration into relatively few documents [12]. Both of these methods are performed as a preprocessing step.

As mentioned, performing variable selection as part of preprocessing can negatively impact the quality of clustering. Additionally, the currently available variable selection methods are mostly focused on Gaussian mixtures. Therefore, variable selection for text and other spherical data in the context of clustering is an important, but relatively unexplored, area of study. In this regard, our proposed method aims to identify irrelevant clustering variables, which can be "noise" or "redundant" variables, by making use of two specialized expectation-maximization algorithms.

This article is organized as follows. Section 2 recalls the the von Mises-Fisher distribution and presents our algorithm for identifying redundant a noise dimensions to perform variable selection. Sections 3 and 4 demonstrates the applicability of our method on simulated and real-world data, respectively. Discussion of these results and concluding remarks are given in Section 5. Detailed derivations are provided in the Appendix (Section A).

2 METHODS

The proposed method seeks to partition the p variables available for clustering into several sets. Using the notation from [14], we aim to find S, the variables relevant for clustering, and S^C, variables which are irrelevant for clustering. S^C is further partitioned into U, the set of redundant variables which can be explained by a variable in R, a subset of S, and W, the set of noise variables, which are not explained by any variables in S. These relations between these variables are presented in Figure 1.

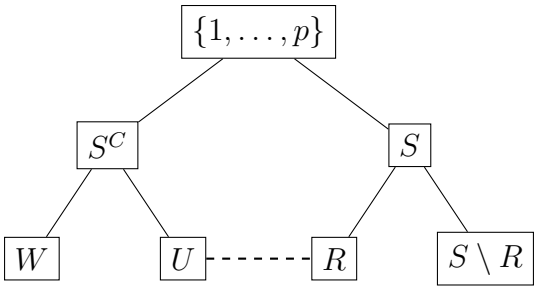


Figure 1: Graphical representation of variable partitioning

To achieve this partitioning, we derive three expectation-maximization (EM) algorithms, which are explained in the following section. First we recall the von Mises-Fisher (vMF) distribution.

A unit random vector $\boldsymbol{x} \in \mathbb{R}^p$ has a p-variate vMF distribution when its probability density function is given by

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \kappa) = c_p(\kappa) e^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}}, \tag{2.1}$$

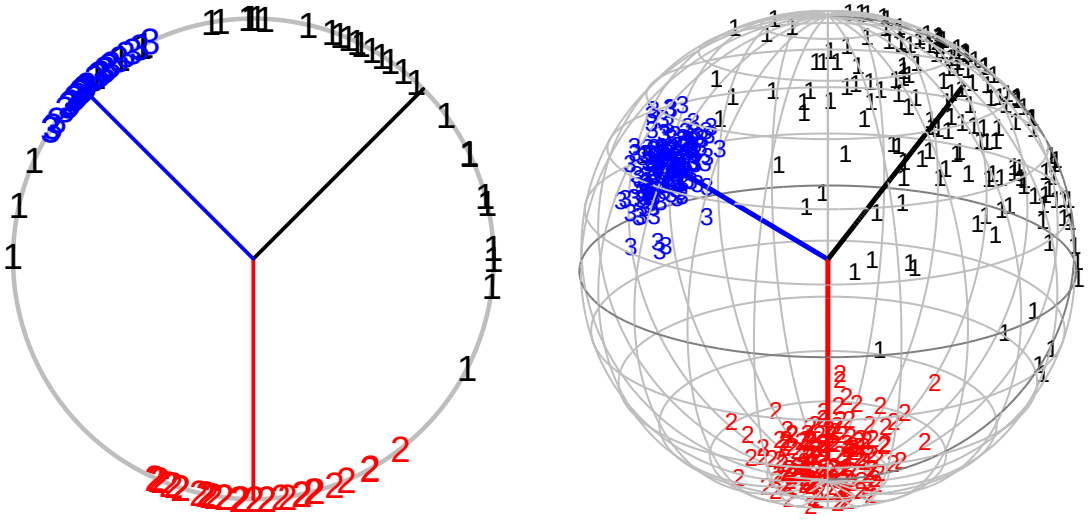
with

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}$$

and $\|\boldsymbol{\mu}\| = 1$, $\kappa \geq 0$, $p \geq 2$, and I_r is the modified Bessel function of the first kind. The mean direction of the distribution is specified by $\boldsymbol{\mu}$. The concentration is given by κ, with $\kappa = 0$ being a uniform density on the hypersphere and $\kappa \to \infty$ approaching a point density at $\boldsymbol{\mu}$. We can then consider a mixture of K vMF distributions indexed by $h \in \{1, \ldots, K\}$, with parameters $\boldsymbol{\theta}_h = (\boldsymbol{\mu}_h, \kappa_h)$, with $\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h = 1$ and mixing proportions α_h, with $\sum_{h=1}^K \alpha_h = 1$. The density of the mixture is given by

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{h=1}^{K} \alpha_h c_p(\kappa_h) \exp\{\kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}\}.$$

Figure 2 shows plots of samples from two mixtures of three vMF distributions with parameters given in Table 1.



(a) 2D (b) 3D

Figure 2: Samples from mixtures of three vMF distributions

Table 1: Parameters used to simulate data in Figure 2

(a) 2D

h	α	κ	μ
1	0.33	2	$\frac{1}{\sqrt{2}}\langle 1, 1\rangle$
2	0.33	20	$\langle 0, -1\rangle$
3	0.34	40	$\frac{1}{\sqrt{2}}\langle -1, 1\rangle$

(b) 3D

h	α	κ	μ
1	0.33	5	$\frac{1}{\sqrt{3}}\langle 1, 1, 1\rangle$
2	0.33	50	$\frac{1}{\sqrt{5}}\langle 0, -2, -1\rangle$
3	0.34	100	$\frac{1}{\sqrt{3}}\langle -1, -1, 1\rangle$

Then the complete data likelihood of a set of observations $\boldsymbol{x}$ is

$$L_c(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{h=1}^{K} \left[\alpha_h c_p(\kappa) \exp\{\kappa_h \boldsymbol{\mu_h}^T \boldsymbol{x}_i\}\right]^{I(z_i=h)} \tag{2.2}$$

where z_i is a latent variable indicating the true component membership of $\boldsymbol{x_i}$. Thus, the complete data log-likelihood is

$$l_c(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{h=1}^{K} I(z_i = h) \left[\ln \alpha_h + \ln c_p(\kappa) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x_i}\right] \tag{2.3}$$

Given some data, we seek to estimate the parameters of this model by maximum likelihood. Because z_i is unknown, we develop an EM algorithm to achieve this goal. Broadly, the EM algorithm consists of an expectation (E) step and a

maximization (M) step. In the E-step, the unknown random variables z_i are estimated based on the current estimates of the parameters. In the M-step, the parameters are updated based on the current estimates of z_i. These steps are iterated until the relative change in the log-likelihood falls below some preset tolerance.

2.1 EM ALGORITHM FOR UNCONSTRAINED VARIABLES

With the constraints $\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h = 1$ and $\kappa_h \geq 0$, the conditional expectation of the complete-data log-likelihood, commonly known as the Q-function, is

$$Q(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{h=1}^{K} \pi_{ih} \ln(\alpha_h) + \sum_{i=1}^{n} \sum_{h=1}^{K} \pi_{ih} \left[\ln(c_p(\kappa_h)) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}_i \right] + \xi \left(1 - \sum_{h=1}^{K} \alpha_h \right) + \sum_{h=1}^{K} \lambda_h (\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h - 1), \quad (2.4)$$

where Lagrangian multipliers ξ and λ_h for $h \in \{1, \ldots, K\}$ and are added for the constraints. In the E-step we obtain the following update equation:

$$\pi_{ih} = \frac{\alpha_h f_h(\boldsymbol{x}_i \mid \boldsymbol{\theta})}{\sum_{h=1}^{K} \alpha_h f_h(\boldsymbol{x}_i \mid \boldsymbol{\theta})} \quad (2.5)$$

In the M-step we obtain the following update equations:

$$\alpha_h = \frac{1}{n} \sum_{i=1}^{n} \pi_{ih} \quad (2.6)$$

$$r_h = \sum_{i=1}^{n} \pi_{ih} \boldsymbol{x}_i \quad (2.7)$$

$$\boldsymbol{\mu}_h = \frac{\boldsymbol{r}_h}{\|\boldsymbol{r}_h\|} \quad (2.8)$$

Derivations for these update equations are provided in Section A.1. An analytical estimate for κ_h is not readily available and is instead estimated by a

numerical optimization procedure. In this case, we use the default procedure from the movMF R package [10], "a variant of the Newton-Fourier method for strictly increasing concave functions." Several other methods for estimating κ_h are available in the package and described in [10]. Henceforth, we refer to these update steps given in Equations 2.5–2.8 as the "Standard-EM" procedure.

2.2 EM ALGORITHM FOR REDUNDANT VARIABLES

In this context, we define redundant variables as those that have a common mean direction within a given mixture component. Let G be set of indices of redundant variables. G can be thought of as a set of candidate variables for U. For each component, h, we denote this common mean by μ_{hG}. We let μ_{hj} be the mean for the jth variable, with $j \notin G$, in the hth component. With this assumption and the constraints $\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h = 1$ and $\kappa_h \geq 0$, we develop the following constrained Q-function:

$$\begin{aligned} Q_{\mathrm{R}}^*(\boldsymbol{\theta} \mid \boldsymbol{x}_i) = & \sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih}\left(\ln(\alpha_h) + \ln(c_p(\kappa_h))\right) \\ & + \sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih}\kappa_h \left(\mu_{hG}\sum_{j\in G} x_{i,j} + \sum_{j\notin G}\mu_{hj}^2 x_{i,j}\right) \\ & + \xi\left(1 - \sum_{h=1}^{K}\alpha_h\right) + \sum_{h=1}^{K}\lambda_h\left[1 - |G|\,\mu_{hG}^2 - \sum_{j\notin G}\mu_{hj}\right] \end{aligned} \tag{2.9}$$

The E-step yields the same update equation as in Equation (2.5). In the M-step, we obtain Equation (2.6), as well as the following update equations:

$$r_{hj} = \sum_{i=1}^{n} \pi_{ih} x_{i,j} \tag{2.10}$$

$$\mu_{hG} = \frac{\sum_{j\in G} r_{hj}}{\sqrt{\left(\frac{1}{|G|}\left(\sum_{j\in G} r_{hj}\right)^2 + \sum_{j\notin G}\left(r_{hj}\right)^2\right)|G|}} \tag{2.11}$$

$$\mu_{hj} = \frac{r_{hj}}{\sqrt{\left(\frac{1}{|G|}\left(\sum_{j\in G} r_{hj}\right)^2 + \sum_{j\notin G}\left(r_{hj}\right)^2\right)}} \tag{2.12}$$

with $|G|$ denoting the number of indices in the set G. Derivations for these update equations are provided in Section A.2. In this case, κ_h is found using the *optimize* function in R [20]. We refer to these update steps as the "Redundant-EM" procedure.

2.3 EM ALGORITHM FOR NOISE VARIABLES

We define a noise variable as those with a common mean direction over all components. Let G be the index of such a variable. G can be thought of as a candidate variable for W. We denote $\mu_{.j}$ to be the common mean of the jth variable over all components. With this assumption and the constraints $\boldsymbol{\mu}_h^T\boldsymbol{\mu_h} = 1$ and $\kappa_h \geq 0$, we obtain the following constrained Q-function:

$$\begin{aligned} Q^*_{\mathrm{N}}(\boldsymbol{\theta} \mid \boldsymbol{x}_i) = &\sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih}\left(\ln(\alpha_h) + \ln(c_p(\kappa_h))\right) \\ &+ \sum_{h=1}^{K}\sum_{i=1}^{n} \pi_{ih}\kappa_h\left(\sum_{j\in G}\mu_{.j}x_{i,j} + \sum_{j\notin G}\mu_{hj}x_{i,j}\right) \\ &+ \xi\left(1 - \sum_{h=1}^{K}\alpha_h\right) + \sum_{h=1}^{K}\lambda_h\left(1 - \sum_{j\in G}\mu_{.j}^2 - \sum_{j\notin G}\mu_{hj}^2\right) \end{aligned} \tag{2.13}$$

In the E-step, Equation (2.5) stays the same. In the M-step, Equation (2.6) stays the same, while the mean directions are updated using the following equations:

$$\mu_{hj} = \frac{\kappa_h r_{hj}}{2\lambda_h} \tag{2.14}$$

$$\mu_{.j} = \frac{\kappa_h \sum_{h=1}^{K} r_{hj}}{2\sum_{h=1}^{K}\lambda_h} \tag{2.15}$$

$$1 = \sum_{j \in G} \mu_{\cdot j}^2 + \sum_{j \notin G} \mu_{hj}^2 \tag{2.16}$$

See Section A.3 for derivations. We note that (2.14) and (2.15) are dependent on λ_h and κ_h. Thus, the closed form solutions for these updates are not available. Because of this, we implement a numerical optimization procedure to approximate solutions. In this case, we use the *optim* function in R [20] with the L-BFGS-B method [5]. We estimate κ_h in the same way as in Section 2.2. We refer to these update steps as the "Noise-EM" procedure.

2.4 VARIABLE SELECTION ALGORITHMS

In this section we propose two methods for identifying redundant and noise variables. The three previously presented EM procedures described in Sections 2.1 - 2.3 are used to determine a soft clustering for a given data X over a mixture of K vMF distributions and a partition of variables into S, R, U, and W sets. For both types of variables, we propose a backward stepwise algorithm which is relatively computationally expensive and conservative as well as a greedy algorithm which can be performed more quickly for on larger data and identifies irrelevant variables more aggressively. The noise and redundant algorithms can be applied in sequence in order to fully partitions the variables. In general, the Bayesian Information Criterion (BIC) [22] is used to determine whether a single variable is noise or if a pair of variables is redundant. Recall the general form of the BIC,

$$\text{BIC} = \ln(n) \cdot C - 2 \cdot l(\hat{\boldsymbol{\theta}} \mid \boldsymbol{x}), \tag{2.17}$$

where C is the number of parameters in the model and $l(\hat{\boldsymbol{\theta}} \mid \boldsymbol{x})$ is the maximized likelihood from Equation (2.3). For a given number of components, K, values for C in each of the models are presented in Table 2.

Table 2: Number of parameters for the three considered models.

Model Type	C
Standard	$K-1+K\times p+K$
Redundant	$K-1+K\times(p-1)+K$
Noise	$K-1+K\times(p-1)+K+1$

2.4.1 Greedy Variable Selection Methods

Below we describe two greedy algorithms (Algorithm 1 and 2) which can be performed with large datasets to identify irrelevant variables aggressively. To identify redundant variables, we begin with all variables in S and find pairs of redundant variables using the BIC criterion. We partition the variables from these pairs to maximize the size of U and minimize the size of R while maintaining that every variable in U is explained by some variable in R. A final model is then fit using only the variables in S. The details are given in Algorithm 1. The greedy algorithm for noise variables (Algorithm 2) is performed similarly. We begin with all variables in S and find noise variables using the BIC criterion. All of these noise variables are appended to U. A final model is fit using only the variables in S. The details are given in Algorithm 2.

Input: Matrix X of data points on p-dimensional unit hypersphere.
Output: Soft clustering of X over a mixture of K vMF distributions; Partition of variables into S, R and U sets.

Step 1

Begin with $S = \{1, \ldots, p\}$, empty list L of variable pairs;
Fit full model, $\mathcal{M}_0$, to X using the Standard-EM procedure;
Obtain $\text{BIC}_{\mathcal{M}_0}$;

Step 2

for *each pair* $i, j \in S$ **do**
    Fit model, $\mathcal{M}_{i,j}$, to X using the Redundant-EM procedure with $G = \{i, j\}$;
    Obtain $\text{BIC}_{\mathcal{M}_{i,j}}$;
    if $BIC_{\mathcal{M}_{i,j}} \leq BIC_{\mathcal{M}_0}$ **then**
        Append the pair i, j to L;
    end
end

Step 3

Let A be the set of unique variables found in L;
for *each pair* $\in L$ **do**
    Append the variable with the least total appearances in L to U. The other variable remains in S;
end
Append $A \setminus U$ to R;
Set $S = \{1, \ldots, p\} \setminus U$;

Step 4

Let X' be the variables in S projected onto the unit hypersphere;
Fit a model, $\mathcal{M}_{\text{R}}$, to X' using the Standard-EM procedure to obtain soft clustering of X;

Algorithm 1: Greedy vMFM for redundant variables

Input: Matrix X of data points on p-dimensional unit hypersphere.
Output: Soft clustering of X over a mixture of K vMF distributions; Partition of variables into S, R, and U sets.
Step 1
Begin with $S = \{1, \ldots, p\}$;
Fit full model, $\mathcal{M}_0$, to X using the Standard-EM procedure;
Obtain $\text{BIC}_{\mathcal{M}_0}$;
Step 2
for *each* $l \in S$ **do**
  Fit model, $\mathcal{M}_l$, to X using the Noise-EM procedure with $G = l$;
  Obtain $\text{BIC}_{\mathcal{M}_l}$;
  if $BIC_{M_l} \leq BIC_{M_0}$ **then**
    Append l to W;
  end
end
Step 3
Set $S = \{1, \ldots, p\} \setminus W$;
Let X' be the variables in S projected onto the unit hypersphere;
Fit a model, $\mathcal{M}_{\text{R}}$, to X' using the Standard-EM procedure to obtain soft clustering of X;

Algorithm 2: Greedy vMFM for noise variables

2.4.2 Backward Stepwise Variable Selection Methods

For smaller datasets we propose more thorough backward stepwise algorithms (Algorithm 3 and 4). To identify redundant variables, we begin with all variables in S and find pairs of redundant variables using the BIC criterion. We then identify the pair which produces the minimum BIC. Two models are then fit to the data, with each excluding one variable in the pair. The variable whose exclusion results in the minimum BIC is appended to U. The other is appended to R. This repeats with the new S set, which excludes the variables in U, until no variable's removal results in a BIC reduction. A final model is fit using only the variables S. Details are given in Algorithm 3. The backward stepwise algorithm for noise variables is performed similarly. We begin with all variables in S and find noise variables using the BIC criterion. The noise variable which results in the lowest BIC is appended to W and removed from S. This repeats with the new S set until no variable's removal results in a BIC reduction. A final model is fit using only the variables S. Details are given in Algorithm 4.

Input: Matrix X of data points on p-dimensional unit hypersphere.
Output: Soft clustering of X over a mixture of K vMF distributions; Partition of variables into S and W sets.

Step 1
Begin with $S = \{1, \ldots, p\}$;
Fit full model, $\mathcal{M}_0$, to X using the Standard-EM procedure;
Obtain $\text{BIC}_{\mathcal{M}_0}$;
Step 2
while $|S| > 2$ **do**
  for *each pair* $i, j \in S$ **do**
    Fit model, $\mathcal{M}_{i,j}$, to X using the Redundant-EM procedure with $G = \{i, j\}$;
    Obtain $\text{BIC}_{\mathcal{M}_{i,j}}$;
  end
  Let i', j' be the minimum $\text{BIC}_{\mathcal{M}_{i,j}}$;
  if $BIC_{\mathcal{M}_{i',j'}} > BIC_{\mathcal{M}_0}$ **then**
    break;
  end
  else
    Let $X_{-i'}$ be the variables $S \cup \{i'\}$ projected onto the unit hypersphere;
    Fit model, $\mathcal{M}_{-i'}$, to $X_{-i'}$ using the Standard-EM procedure;
    Obtain $\text{BIC}_{\mathcal{M}_{-i'}}$;
    Let $X_{-j'}$ be $S \cup \{j'\}$ projected onto the unit hypersphere;
    Fit model, $\mathcal{M}_{-j'}$, to $X_{-j'}$ using the Standard-EM procedure;
    if $BIC_{\mathcal{M}_{-i'}} < BIC_{\mathcal{M}_{-j'}}$ **then**
      Let $X = X_{-i'}$;
      Append i' to U;
      Set $S = S \setminus \{i'\}$;
      Append j' to R;
    end
    else
      Let $X = X_{-j'}$;
      Append j' to U;
      Set $S = S \setminus \{j'\}$;
      Append i' to R;
    end
  end
end
Step 3
Let X' be the variables S projected onto the unit hypersphere;
Fit a model, $\mathcal{M}_{\text{R}}$, to X' using the Standard-EM procedure to obtain soft clustering of X;

Algorithm 3: Stepwise vMFM for redundant variables

Input: Matrix X of data points on p-dimensional unit hypersphere.
Output: Soft clustering of X over a mixture of K vMF distributions; Partition of variables into S and W sets.
Step 1
Begin with $S = \{1, \ldots, p\}$;
Fit full model, $\mathcal{M}_0$, to X using the Standard-EM procedure;
Step 2
while $|S| > 2$ **do**
  for *each* $l \in S$ **do**
    Fit model, $\mathcal{M}_l$, to X using the Noise-EM procedure with $G = l$;
    Obtain $BIC_{\mathcal{M}_l}$;
  end
  Let l' be the minimum $\text{BIC}_{\mathcal{M}_l}$;
  if $BIC_{\mathcal{M}_{l'}} > BIC_{\mathcal{M}_0}$ **then**
    break;
  end
  else
    Append l' to W;
    Set $S = S \setminus \{l'\}$;
    Let X be the variables S projected onto the unit hypersphere.;
  end
end
Step 3
Let X' be the variables S projected onto the unit hypersphere;
Fit a model, $\mathcal{M}_{\text{R}}$, to X' using the Standard-EM procedure to obtain soft clustering of X;

Algorithm 4: Stepwise vMFM for noise variables

2.5 COMPUTATIONAL ASPECTS

Initialization is performed using the *em-EM* method [2]. In this method, we start by randomly choosing K points as the means for our K vMF distributions and initializing z_i as the closest mean to one of the random seeds using cosine distance. We run a *short-EM* algorithm by iterating until the relative change in log-likelihood is less than a lax tolerance level. This procedure is repeated a fixed number of times and the parameter estimates with the highest likelihood value are used as initial points to run the *long-EM* algorithm until a more strict convergence criterion is met. In the simulation study (Section 3) and the applications (Section 4), where $K \leq 3$, 10 random seeds are generated and a tolerance level of 10^{-2} is used for the *short-EM*. The *long-EM* is run using

tolerance level of 10^{-6}. For applications with larger K, the number of random seeds should be increased or an alternative initialization method, such as the one proposed in [18], should be used. Additionally, κ is restricted to be less than 1500 when performing numerical optimization.

We note the greedy redundant algorithm is $O(p^2(n+p)k)$ and the greedy noise algorithm is $O(p(n+p)k)$. In the worst-case, the stepwise redundant algorithm is $O(p^3(n+p)k)$ and the stepwise noise algorithm is $O(p^2(n+p)k)$. In practice, fitting these complex models for all possible pairs of redundant variables and all possible noise variables may be too computationally expensive. In this case, we recommend only fitting models for pairs of variables which are likely to be redundant and variables which are likely to be noise. To identify likely pairs of redundant variables, after fitting M_0, we compute the Euclidean distance between the means of each variable over all components. The m pairs with the smallest distance are identified as likely redundant pairs and tested using the model fitting procedure described above. To identify likely noise variables, after fitting M_0, we compute the variance of each variable's mean over all components. The m pairs with the smallest distance are identified as likely noise variables and tested using the model fitting procedure described above.

3 SIMULATION STUDY

An extensive study is conducted to evaluate the performance of the proposed method and algorithm using simulated data. A 2-component, 8-dimensional mixtures of vMF distributions was generated. From this mixture $N = 1000$ datasets of size $n = 1000$ observations are generated using the *rmovMF* function from the *movMF* package [10]. Parameters for the mixture are given in Table 3. The variables are created so that there will be some pairs which are redundant and some variables which are noise. From the $\boldsymbol{\mu}_h$ column in Table 3, we note the following partitioning of variables: $W = (4, 5)$, $U = (2, 3)$, $R = (1)$, $S =$

$(1, 6, 7, 8)$. The means and concentration parameters were chosen to make the mixture easily separable while maintaining the redundant and noise variable relationships stated above.

Table 3: Parameters of simulated data

h	α_h	κ_h	μ_h
1	0.4	4	$\frac{1}{\sqrt{38}}\langle -3, -3, -3, -1, 1, 0, 0, 3\rangle$
2	0.6	4	$\frac{1}{\sqrt{70}}\langle 3, 3, 3, -1, 1, 4, -4, -3\rangle$

We applied the greedy vMFM for redundant and for noise variables to these simulated datasets. The step-by-step values of the BIC and relative differences for one example dataset are presented in Tables 4 and 5. $\mathcal{M}_{1,2}$, $\mathcal{M}_{1,3}$, and $\mathcal{M}_{2,3}$ are models corresponding to pairs of redundant variables and $\mathcal{M}_4$ and $\mathcal{M}_5$ correspond to noise variables. In the Tables 4 and 5, RelDiff(%) indicates the relative difference in BIC between the base model, $\mathcal{M}_0$ or $\mathcal{M}_{\mathrm{R}}$, and the model in consideration $\mathcal{M}_{i,j}$. This is given by $\frac{\mathrm{BIC}_{\mathcal{M}_{i,j}} - \mathrm{BIC}_{\mathcal{M}_0}}{\mathrm{BIC}_{\mathcal{M}_0}}$. A positive relative difference in BIC indicates an improvement from the base model, while a negative relative difference in BIC indicates the considered model is inferior to the base model. From Table 4, we note a slight decrease in BIC for pairs of variables which were simulated to be redundant and larger increase in BIC for variables which were simulated to not be redundant. Similarly, from Table 5, we note a slight decrease in BIC for variables which were simulated to be noise and a larger increase in BIC for variables which were simulated to not be noise. In the case of our methods being applied to variables which were simulated to be redundant or noise, the improved BIC arises because we are able to achieve similar likelihood values by reducing the parameter counts. In the case of our methods being applied to other variables, the increased BIC is a result of the substantially lower likelihood of the data under the model, which is not outweighed by the reduced parameter count. In 1000 datasets, at least two pairs of redundant were correctly identified in all simulations, while all three were correctly identified in 999 of the simulations (99.9%). With regard to noise

variables, out of 1000 simulations, at least one noise column were correctly identified in 999 simulations (99.9%), while both noise columns were correctly identified in 979 simulations (97.9%).

Table 4: Example results from simulated data

Model	BIC	RelDiff(%)	$l(\hat{\boldsymbol{\theta}} \mid \boldsymbol{x})$	C
$\mathcal{M}_0$	-582.20	0.00	356.72	19
$\mathcal{M}_{1,2}$	-595.30	2.25	356.37	17
$\mathcal{M}_{1,3}$	-595.38	2.26	356.41	17
$\mathcal{M}_{2,3}$	-593.32	1.91	355.38	17
$\mathcal{M}_{6,7}$	-257.03	-55.85	187.23	17

Table 5: Noise results from simulated data

Model	BIC	RelDiff(%)	$l(\hat{\boldsymbol{\theta}} \mid \boldsymbol{x})$	C
$\mathcal{M}_{\mathrm{R}}$	-376.07	0.00	239.84	15
$\mathcal{M}_4$	-380.94	1.29	238.82	14
$\mathcal{M}_5$	-377.78	0.46	237.25	14
$\mathcal{M}_8$	-276.27	-26.54	186.49	14

4 APPLICATION

In this section we assess our methods on a variety of real-world datasets. When computationally feasible we apply the stepwise vMFM algorithms to identify irrelevant variables. For larger datasets, we apply the greedy vMFM algorithms.

4.1 CLASSIC3 DATA

We evaluated our algorithm using the well-known Classic3 document collection[1]. The collection consists of 3891 total documents, including 1398 Cranfield (cran) documents from aeronautical system papers, 1033 Medline (med) documents from medical journals, and 1460 CISI (cisi) documents from information retrieval papers ($K = 3$). This data contained 21,137 unique words, meaning that each document was represented as a 21,137-dimensional vector.

[1] ftp://ftp.cs.cornell.edu/pub/smart

In addition, balanced and unbalanced subsets of the data, which were chosen by randomly selecting a fixed number of documents from each group, were also considered. This subset sampling was repeated 1000 times. Table 6 details the composition and dimensionality of each dataset. As expected, the dimensionality of the sampled datasets are smaller than the original.

Table 6: Descriptions of the classic3 datasets

Name	cisi	cran	med	Avg. p
Classic3	1460	1398	1033	21137.00
Classic3 300 Balanced	100	100	100	6090.47
Classic3 400 Unbalanced	100	200	100	6671.66

Several variable selection methods specific to text data are considered in addition to our vMFM methods. The methods are described below.

- none: no variable selection performed
- none+red+m+dist: m pairs of potential redundant variables (selected by smallest euclidean distance between means) considered with greedy vMFM for redundant variables.
- stop: all stop words in [23] removed.
- stop+red+m+dist: all stop words in [23] removed and n pairs of potential redundant variables (selected by smallest euclidean distance between means) considered with greedy vMFM for redundant variables.
- tf-idf: the same number of words removed as with none+red+m+dist by lowest term frequencyinverse document frequency.
- stop+tf-idf: all stop words in [23] removed and the same number of words removed as with stop+red+m+dist by lowest term frequencyinverse document frequency.

Tables 7–9 present results. Because tf-idf and none+red+m+dist remove the same number of dimensions, their choice of dimensions to remove can be directly compared. Similarly, stop+tf-idf and stop+red+m+dist remove the same

number of dimensions, so they can be compared directly. All variable selection techniques successfully improved classification rates on the full dataset. While the greedy vMFM for redundant variables was able to improve or maintain accuracy, we note that removing the same number of dimensions by tf-idf resulted in an additional improvement in accuracy by about 10%.

Table 7: Results of various variable selection methodologies on the Classic3 dataset

Processing	Accuracy (%)	p Removed
stop+tf-idf	69.49	975
stop	68.90	618
stop+red+1000+dist	60.55	975
tf-idf	55.38	1000
none+red+1000+dist	44.02	1000
none	43.95	0

Table 8: Averaged results of various variable selection methodologies on balanced subsets of 300 documents from the Classic3 data set

Processing	Avg. Accuracy (%)	Avg. Rank	Avg. p Removed
stop+tf-idf	77.40	2.00	654.30
stop+red+1000+dist	76.34	2.08	654.30
stop	76.13	2.17	487.24
tf-idf	57.69	3.80	163.43
none	42.47	5.28	0.00
none+red+1000+dist	42.23	5.67	163.43

Table 9: Averaged results of various variable selection methodologies on unbalanced subsets of 400 documents from the Classic3 data set

Processing	Avg. Accuracy (%)	Avg. Rank	Avg. p Removed
stop	61.80	2.24	501.04
stop+red+1000+dist	60.96	2.40	638.44
stop+tf-idf	59.76	2.44	638.44
tf-idf	54.92	3.00	135.82
none	42.43	5.35	0.00
none+red+1000+dist	42.44	5.57	135.82

In both subsets, we again note that methods which remove all stop words perform much better than methods that do not. The three stop word methods

all perform similarly, with a maximum difference in accuracy of about 2%. Reducing dimensions using the greedy vMFM algorithm for redundant variables after removing stop words enables greater variable selection beyond removing stop words alone, without necessarily sacrificing performance. Similar performance to removing stop words and words by tf-idf, but we hypothesize that all three methods could be combined to further reduce dimensionality.

Methods which did not involve the removal of stop words performed significantly worse. In both the unbalanced and balanced datasets, applying the greedy vMFM for redundant variables led to the slightly lower average classification accuracy (by 0.01% or 0.024%) than performing no variable selection at all and enabled a significant reduction in variables involved in the model. In both cases, removing word by tf-idf improved classification rates while reducing dimensionality.

4.2 BREAST CANCER WISCONSIN DATA

We also evaluate our stepwise methods on the Wisconsin Breast Cancer dataset, obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [26]. The data reports 9 discrete measurements for 699 observations of clumps of breast cancer cells. The variables are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. A correlation matrix for these variables is presented in Table 10. We remove 16 missing observations from the dataset to arrive at a final dataset with $n = 683$, $p = 9$. Model-based clustering alanysis is conducted using the proposed mixture model based on these variables in order to predict whether a given clump belongs to the class benign or malignant.

Table 10: Within class correlation between variables in the Wisconsin Breast Cancer dataset (Benign/Malignant)

	Clump Thickness	Uniformity Cell Size	Uniformity Cell Shape	Marginal Adhesion	Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
Clump Thickness	1/1								
Uniformity Cell Size	.28/.1	1/1							
Uniformity Cell Shape	.3/.11	.7/.72	1/1						
Marginal Adhesion	.26/-.14	.28/.32	.24/.27	1/1					
Epithelial Cell Size	.16/.02	.41/.46	.34/.38	.29/.19	1/1				
Bare Nuclei	.12/-.04	.46/-.04	.36/.05	.37/.19	.33/-.03	1/1			
Bland Chromatin	.1/-.02	.26/.39	.19/.34	.12/.34	.15/.22	.21/.14	1/1		
Normal Nucleoli	.21/-.01	.49/.3	.39/.31	.25/.18	.44/.23	.31/-.08	.34/.25	1/1	
Mitoses	-.04/.12	.05/.24	0/.21	.06/.2	-.02/.33	.12/-.04	-.04/.06	.06/.22	1/1

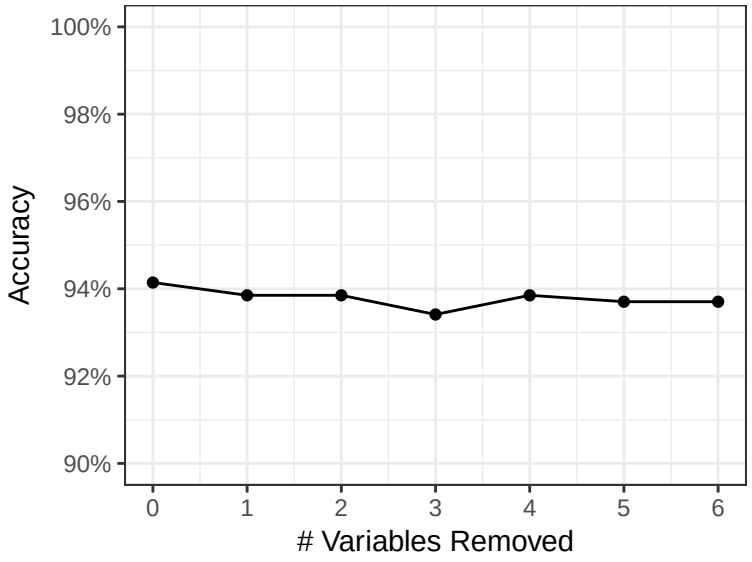


Figure 3: Wisconsin Breast Cancer accuracy after removing redundant variables with stepwise procedure

Using the whole dataset, we achieve a peak classification rate of 94.14%. Using the stepwise vMFM for noise variables, we do not identify any variables as noise. As shown in Figure 3, the stepwise vMFM for redundant variables results in the removal of six of the nine variables, with only Uniformity of Cell Size, Bland Chromatin, and Mitoses remaining. Using only these three variables reduces accuracy from 94.14% to 93.70%. Looking more closely at the result of each step of the stepwise algorithm, we see that Clump Thickness was removed because of its redundancy with Uniformity of Cell Shape, while all of the other variables removed, including Uniformity of Cell Shape, were found to be redundant with Uniformity of Cell Size. These relationships generally make sense, given the information presented in Table 10, where we note the strong correlations between Uniformity of Cell Size and the other removed variables. We also note that the other remaining variables, Mitoses and Bland Chromatin have relatively low correlation with the removed variables.

5 DISCUSSION

Our greedy vMFM for noise and redundant variables algorithms were shown to perform well on simulated data. Applying our methods to real data yielded mixed results. In the best cases, we were able to significantly reduce dimensionality without sacrificing much accuracy. In the case of text data, our methods were shown to perform well when combined with another standard dimension reduction technique, the removal of stop words, but were not shown to consistently outperform another standard technique, the removal of words with low term frequency inverse document frequency. Areas for future development include the adoption of more relaxed linear relationships between redundant variables like those presented in [14], a more thorough examination of methods for selecting likely redundant or noise variables, and addressing computational concerns which prevented us from identifying noise variables in larger datasets.

A APPENDIX

A.1 DERIVATIONS FOR UNCONSTRAINED VARIABLES

A.1.1 E-step

In the E-step of our algorithms, we update the distribution estimates of the hidden variables, z_i, by evaluating the expectation of Equation (2.3) assuming all other parameters are known:

$$\begin{aligned}\mathbb{E}\left(l_c(\boldsymbol{\theta} \mid \boldsymbol{x})\right) &= \mathbb{E}\left(\sum_{i=1}^{n}\sum_{h=1}^{K} I(z_i = h)\left[\ln(\alpha_h) + \ln(c_p(\kappa)) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}\right]\right) \\ &= \sum_{i=1}^{n}\sum_{h=1}^{K} \mathbb{E}(I(z_i = h))\left[\ln(\alpha_h) + \ln(c_p(\kappa)) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x_i}\right] \\ &= \sum_{i=1}^{n}\sum_{h=1}^{K} P(h \mid \boldsymbol{x}_i, \boldsymbol{\theta})\left[\ln(\alpha_h) + \ln(c_p(\kappa)) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x_i}\right] \qquad \text{(A.1)}\end{aligned}$$

This expectation reduces to calculating the posterior probability of the component h given observation $\boldsymbol{x}_i$ and parameter values, $P(h \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ which we denote as π_{ih}.

$$\begin{aligned}\pi_{ih} =& P(h \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ =& \frac{P(h, \boldsymbol{x}_i, \boldsymbol{\theta})}{P(\boldsymbol{x}_i, \boldsymbol{\theta})} \\ =& \frac{P(\boldsymbol{x}_i, \boldsymbol{\theta} \mid h)P(h)}{P(\boldsymbol{x}_i, \boldsymbol{\theta})} \\ =& \frac{\alpha_h f_h(\boldsymbol{x}_i \mid \boldsymbol{\theta})}{\sum_{h=1}^{K} \alpha_h f_h(\boldsymbol{x}_i \mid \boldsymbol{\theta})}\end{aligned}$$

A.1.2 M-step

Next we present three variations of the M-step, one each for the Standard, Redundant, and Noise procedures. In the M-step, we fix π_{ih} and maximize an objective function (Q) which is the expectation of the complete-data log-

likelihood (Equation A.1):

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{x}) &= \sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih} \left[\ln(\alpha_h) + \ln(c_p(\kappa)) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}_i\right] \\ &= \sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih} \ln(\alpha_h) + \sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih} \left[\ln(c_p(\kappa_h)) + \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}_i\right] \end{aligned} \tag{A.2}$$

We note that the first term in the sum can be maximized separately from the second. The form of each α_h is the same for the standard, redundant, and noise cases, so we derive it only once. We now estimate each α_h using

$$Q^*_{\alpha_h}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n}\sum_{h=1}^{K} \pi_{ih} \ln(\alpha_h) + \xi \left(1 - \sum_{h=1}^{K} \alpha_h\right) \tag{A.3}$$

where we have included a Lagrangian multiplier corresponding to the constraint $\sum_{h=1}^{K} \alpha_h = 1$. To maximize Equation (A.3), we differentiate with respect to ξ and each α_h and simplify with the partial derivative being equal to 0. From $0 = \frac{\partial Q^*_{\alpha_h}}{\partial \xi}$, we derive

$$1 = \sum_{h=1}^{K} \alpha_h. \tag{A.4}$$

From $0 = \frac{\partial Q^*_{\alpha_h}}{\partial \alpha_h}$, we derive

$$\xi = -n \tag{A.5}$$

and

$$\alpha_h = \frac{1}{n} \sum_{i=1}^{n} \pi_{ih}. \tag{A.6}$$

We now maximize the second term in Equation (A.2). An analytical estimate for κ_h is not readily available and is instead estimated by a numerical optimization procedure. Because of this, we modify the second term in Equation (A.2), ignoring $c_p(\kappa)$ and introducing a Lagrangian multiplier λ_h corre-

sponding to the constraint $\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h = 1$:

$$Q^*(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{h=1}^{K} \pi_{ih} \kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}_i + \sum_{h=1}^{K} \lambda_h (\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h - 1) \tag{A.7}$$

To maximize Equation (A.7), we differentiate with respect to each variable and simplify with the partial derivative being equal to 0. From $0 = \frac{\partial Q^*}{d\boldsymbol{\mu}_h}$, we derive

$$\boldsymbol{\mu}_h = \frac{\kappa_h}{2\lambda_h} \sum_{i=1}^{n} \pi_{ih} \boldsymbol{x}_i. \tag{A.8}$$

From $0 = \frac{\partial Q^*}{d\lambda_h}$, we derive

$$\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h = 1. \tag{A.9}$$

Using Equation (A.8) and Equation (A.9), we derive

$$\lambda_h = \frac{\kappa_h}{2} \left\| \sum_{i=1}^{n} \pi_{ih} \boldsymbol{x}_i \right\| \tag{A.10}$$

and

$$\boldsymbol{\mu}_h = \frac{\sum_{i=1}^{n} \pi_{ih} \boldsymbol{x}_i}{\left\| \sum_{i=1}^{n} \pi_{ih} \boldsymbol{x}_i \right\|}. \tag{A.11}$$

Letting $\boldsymbol{r}_h = \sum_{i=1}^{n} \pi_{ih} \boldsymbol{x}_i$, we have:

$$\lambda_h = \frac{\kappa_h}{2} \|r_h\| \tag{A.12}$$

and

$$\boldsymbol{\mu}_h = \frac{\boldsymbol{r}_h}{\|\boldsymbol{r}_h\|}. \tag{A.13}$$

A.2 DERIVATIONS FOR REDUNDANT VARIABLES

For the redundant and noise maximizations, we begin by rewriting Equation (A.7) using only scalar multiplication:

$$Q^*(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n}\sum_{h=1}^{K}\pi_{ih}\kappa_h\sum_{j=1}^{p}\mu_{hj}x_{ij} + \sum_{h=1}^{K}\lambda_h\left(\sum_{j=1}^{p}\mu_{hj}^2 - 1\right) \tag{A.14}$$

We define redundant variables to be variables which have a common mean direction within a given mixture component. Let G be set of indices of redundant variables. G can be thought of as a set of candidate variables for U. We can denote this common mean by μ_{hG} for each component h. We let μ_{hj} be the mean for the jth variable in the hth component. With this assumption and the constraints $\boldsymbol{\mu}_h^T\boldsymbol{\mu_h} = 1$ and $\kappa_h \geq 0$, we can rewrite some parts of Equation (A.14) using this new notation:

$$Q_{\mathrm{R}}^*(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n}\sum_{h=1}^{K}\pi_{ih}\kappa_h\left(\mu_{hG}\sum_{j\in G}x_{ij} + \sum_{j\notin G}\mu_{hj}x_{ij}\right) + \sum_{h=1}^{K}\lambda_h\left[1 - |G|\,\mu_{hG}^2 - \sum_{j\notin G}\mu_{hj}\right] \tag{A.15}$$

with $|G|$ denoting the cardinality of G. To maximize Equation (A.15), we differentiate with respect to each variable and simplify with the partial derivative being equal to 0. From $0 = \frac{\partial Q_{\mathrm{R}}^*}{\partial \lambda_h}$, we derive

$$1 = |G|\,\mu_{hG}^2 + \sum_{j\notin G}\mu_{hj}^2. \tag{A.16}$$

From $0 = \frac{\partial Q_{\mathrm{R}}^*}{\partial \boldsymbol{\mu}_h}$, we derive

$$\mu_{hG} = \frac{\kappa_h\sum_{j\in G}\sum_{i=1}^{n}\pi_{ih}x_{ij}}{2\lambda_h|G|}. \tag{A.17}$$

From $0 = \frac{\partial Q_{\mathrm{R}}^*}{\partial \mu_{hj}}$, we derive

$$\mu_{hj} = \frac{\kappa_h\sum_{i=1}^{n}\pi_{ih}x_{ij}}{2\lambda_h}. \tag{A.18}$$

for $j \notin G$. Substituting Equation (A.17) and Equation (A.18) into Equation (A.16) yields

$$\lambda_h = \frac{\kappa_h}{2} \sqrt{\left(\frac{1}{|G|} \left(\sum_{j \in G} \sum_{i=1}^{n} \pi_{ih} x_{ij} \right)^2 + \sum_{j \notin G} \left(\sum_{i=1}^{n} \pi_{ih} x_{ij} \right)^2 \right)}. \tag{A.19}$$

Substituting Equation (A.19) into Equation (A.17) yields

$$\mu_{hG} = \frac{\sum_{j \in G} \sum_{i=1}^{n} \pi_{ih} x_{ij}}{\sqrt{\left(\frac{1}{|G|} \left(\sum_{j \in G} \sum_{i=1}^{n} \pi_{ih} x_{ij} \right)^2 + \sum_{j \notin G} \left(\sum_{i=1}^{n} \pi_{ih} x_{ij} \right)^2 \right) |G|}}. \tag{A.20}$$

Substituting Equation (A.19) into Equation (A.18) yields

$$\mu_{hj} = \frac{\sum_{i=1}^{n} \pi_{ih} x_{ij}}{\sqrt{\left(\frac{1}{|G|} \left(\sum_{j \in G} \sum_{i=1}^{n} \pi_{ih} x_{ij} \right)^2 + \sum_{j \notin G} \left(\sum_{i=1}^{n} \pi_{ih} x_{ij} \right)^2 \right)}}. \tag{A.21}$$

Letting $r_{hj} = \sum_{i=1}^{n} \pi_{ih} x_{ij}$, we have:

$$\mu_{hG} = \frac{\sum_{j \in G} r_{hj}}{\sqrt{\left(\frac{1}{|G|} \left(\sum_{j \in G} r_{hj} \right)^2 + \sum_{j \notin G} \left(r_{hj} \right)^2 \right) |G|}} \tag{A.22}$$

and

$$\mu_{hj} = \frac{r_{hj}}{\sqrt{\left(\frac{1}{|G|} \left(\sum_{j \in G} r_{hj} \right)^2 + \sum_{j \notin G} \left(r_{hj} \right)^2 \right)}}. \tag{A.23}$$

A.3 DERIVATIONS FOR NOISE VARIABLES

We define a noise variable as variable with a common mean direction over all components. Let G be the index of such a variable. G can be thought of as a set of candidate variables for W. We denote $\mu_{.j}$ to be the common mean of the jth variable over all components and μ_{hj} to be the mean of the jth variable (which is not in G) in the hth component. With this assumption and the constraints

$\boldsymbol{\mu}_h^T \boldsymbol{\mu}_{\boldsymbol{h}} = 1$ and $\kappa_h \geq 0$, we can rewrite some parts of Equation (A.14) using this new information:

$$Q_{\mathrm{N}}^*(\boldsymbol{\theta} \mid \boldsymbol{x}_i) = \sum_{h=1}^{K} \sum_{i=1}^{n} \pi_{ih} \kappa_h \left(\sum_{j \in G} \mu_{\cdot j} x_{ij} + \sum_{j \notin G} \mu_{hj} x_{ij} \right) + \sum_{h=1}^{K} \lambda_h \left(1 - \sum_{j \in G} \mu_{\cdot j}^2 - \sum_{j \notin G} \mu_{hj}^2 \right) \tag{A.24}$$

To maximize Equation (A.24), we differentiate with respect to each variable and simplify with the partial derivative being equal to 0. From $0 = \frac{\partial Q_{\mathrm{N}}^*}{\partial \mu_{hj}}$, we derive

$$\mu_{hj} = \frac{\kappa_h r_{hj}}{2\lambda_h}. \tag{A.25}$$

From $0 = \frac{\partial Q_{\mathrm{N}}^*}{\partial \mu_{\cdot j}}$, we derive

$$\mu_{\cdot j} = \frac{\kappa_h \sum_{h=1}^{K} r_{hj}}{2 \sum_{h=1}^{K} \lambda_h}. \tag{A.26}$$

From $0 = \frac{\partial Q_{\mathrm{N}}^*}{\partial \lambda_h}$, we derive

$$1 = \sum_{j \in G} \mu_{\cdot j}^2 + \sum_{j \notin G} \mu_{hj}^2. \tag{A.27}$$

REFERENCES

[1] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

[2] Christophe Biernacki, Gilles Celeux, and Grard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003. ISSN 0167-9473. doi:10.1016/S0167-9473(02)00163-9.

[3] Avleen S. Bijral, Markus Breitenbach, and Greg Grudic. Mixture of watson distributions: A generative model for hyperspherical embeddings. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 35–42, San Juan, Puerto Rico, Mar 2007. PMLR.

[4] Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014. ISSN 0167-9473. doi:https://doi.org/10.1016/j.csda.2012.12.008.

[5] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi:10.1137/0916069.

[6] Gilles Celeux, Cathy Maugis-Rabusseau, and Mohammed Sedki. Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Advances in Data Analysis and Classification*, Apr 2018. ISSN 1862-5355. doi:10.1007/s11634-018-0322-5.

[7] Jean-Luc Dortet-Bernadet and Nicolas Wicker. Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics*, 9(1):66–80, 2008. doi:10.1093/biostatistics/kxm012.

[8] Mojgan Golzy, Marianthi Markatou, and Arti Shivram. Algorithms for clustering on the sphere: Advances & applications. In *Proceedings of The World Congress on Engineering and Computer Science 2016*, volume 1, pages 420–425, 2016.

[9] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Pairwise variable selection for highdimensional modelbased clustering. *Biometrics*, 66(3):793–804, 2010. doi:10.1111/j.1541-0420.2009.01341.x.

[10] Kurt Hornik and Bettina Grn. movmf: An r package for fitting mixtures of von mises-fisher distributions. *Journal of Statistical Software, Articles*, 58(10):1–31, 2014. ISSN 1548-7660. doi:10.18637/jss.v058.i10.

[11] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, Sept 2004. ISSN 0162-8828. doi:10.1109/TPAMI.2004.71.

[12] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, second edition, 2014.

[13] Ranjan Maitra and Ivan P. Ramler. A k-mean-directions algorithm for fast clustering of data on the sphere. *Journal of Computational and Graphical Statistics*, 19(2):377–396, 2010. doi:10.1198/jcgs.2009.08155.

[14] C. Maugis, G. Celeux, and M. L. Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882, 2009. ISSN 0167-9473. doi:10.1016/j.csda.2009.04.013.

[15] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1):9–29, Jan 2001. ISSN 1573-0565. doi:10.1023/A:1007648401407.

[16] Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010. doi:10.1214/09-SS053.

[17] Volodymyr Melnykov, Semhar Michael, and Igor Melnykov. *Recent Developments in Model-Based Clustering with Applications*, pages 1–39. Springer International Publishing, 2015. ISBN 978-3-319-09259-1. doi:10.1007/978-3-319-09259-1_1.

[18] Semhar Michael and Volodymyr Melnykov. An effective strategy for initializing the em algorithm in finite mixture models. *Advances in Data Analysis and Classification*, 10(4):563–583, Dec 2016. ISSN 1862-5355. doi:10.1007/s11634-016-0264-8.

[19] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8: 1145–1164, 2007.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[21] Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006. doi:10.1198/016214506000000113.

[22] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6 (2):461–464, 03 1978. doi:10.1214/aos/1176344136.

[23] Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, 1(3), 2016. doi:10.21105/joss.00037.

[24] Shivakumar Vaithyanathan and Byron Dom. Model-based hierarchical clustering. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI '00, pages 599–608, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-709-9.

[25] Sijian Wang and Ji Zhu. Variable selection for modelbased highdimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448, 2008. doi:10.1111/j.1541-0420.2007.00922.x.

[26] William H Wolberg and Olvi L Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23):9193–9196, 1990. ISSN 0027-8424. doi:10.1073/pnas.87.23.9193.

[27] Benhuai Xie, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, 2:168–212, 2008. doi:10.1214/08-EJS194.

[28] Shi Zhong. Efficient online spherical k-means clustering. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 5, pages 3180–3185, July 2005. doi:10.1109/IJCNN.2005.1556436.

[29] Shi Zhong and Joydeep Ghosh. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3): 374–384, Sep 2005. doi:10.1007/s10115-004-0194-1.

[30] Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Statist.*, 3:1473–1496, 2009. doi:10.1214/09-EJS487.