Variable Selection Techniques for Clustering on the Unit Hypersphere

Semhar Michael

Assistant Professor
Mathematics and Statistics
Jerome Lohr College of Engineering
South Dakota State University

Joint Statistical Meeting

July 30, 2019

This is joint work with Damon Bayer
Acknowledgments

Financial support from Arnold K. Skeie E-Commerce Analytics Faculty and Student Fellowship
Outline

1. Introduction
   - Finite mixture models
   - vMF mixture
   - Parameter estimation

2. Variable selection for clustering
   - Modified EM
   - Variable Selection Algorithm

3. Experiment

4. Application
Clustering on the hypersphere
- Generative (mixtures of von Mises-Fisher distributions) such as Banerjee et al. 2005 [1]
- Non-generative (spherical $K$-means) by Maitra and Ramler, 2010 [8]

Variable selection for model-based clustering
- Primarily for mixtures of normals
- Explicit assumptions relating relevant and irrelevant variables e.g. Maugis et al. 2009 [9]
- Regularization methods by Pan and Shen, 2010 and others [10, 12, 5]
Literature review

- Clustering on the hypersphere
  - Generative (mixtures of von Mises-Fisher distributions) [1, 2, 3, 4]
  - Non-generative (spherical $K$-means) [8, 13]
- Variable selection for model-based clustering
  - Primarily for mixtures of normals
  - Explicit assumptions relating relevant and irrelevant variables [6, 11, 9]
  - Regularization methods [10, 12, 5]
Finite mixture models

\[ g(y; \Psi) = \sum_{h=1}^{K} \alpha_h f_h(y; \vartheta_h) \text{ with } 0 < \alpha_h \leq 1, \sum_{h=1}^{K} \alpha_h = 1 \]

- There are various parameter estimation procedures
- MLE is the most popular
- Standard tool - The Expectation Maximization (EM) algorithm
The Expectation Maximization (EM) algorithm

- The estimation of parameters is done via EM algorithm.
- Introduce a missing information \( Z_{ih} \) where

\[
Z_{ih} = \begin{cases} 
1 & \text{if } y_i \in h\text{th component} \\
0 & \text{otherwise}
\end{cases}
\]

- The complete-data likelihood function can be written as

\[
L_c(\Psi; y) = \prod_{i=1}^{n} \prod_{h=1}^{K} [\alpha_h f_h(y_i; \vartheta_h)]^{Z_{ih}},
\]

- The complete-data log-likelihood function has the form

\[
\ell_c(\Psi; y) = \sum_{i=1}^{n} \sum_{h=1}^{K} Z_{ih} \log [\alpha_h f_h(y_i; \vartheta_h)],
\]
EM algorithm

The Expectation-Maximization (EM) algorithm: two steps

- **E-step:**
  - Calculates posterior probabilities, \( \pi_{ih} = P(Z_{ih} = h) \) (i.e. the posterior probability that the observation \( y_i \) came from the \( h \)th mixture component)

- **M-Step:**
  - Maximizes the conditional expectation of the complete-data log-likelihood function given observed data

\[
Q(\Psi; \Psi^{(b-1)}, y) = \sum_{i=1}^{n} \sum_{h=1}^{K} \pi_{ih}^{(b)} \left( \log \alpha_h + \log f_h(y_i; \vartheta_h) \right)
\]

- Iterate until a convergence criterion is met
Mixture of von-Mises Fisher distribution

- vMF distribution is defined for a $p$-dimensional unit random vector $x$, with restriction $\|x\| = 1$
- The density is given as

$$f(x|\vartheta) = c_p(\eta) \exp \{ \eta \mu^T x \},$$

where $\vartheta = \{\mu, \eta\}$ are the mean direction and concentration parameters respectively. $c_p(\eta) = \eta^{p/2 - 1} / \left\{ (2\pi)^{p/2} I_{p/2 - 1}(\eta) \right\}$ and $I_r(.)$ is an order $r$ modified Bessel function of first kind and $\|\mu\| = 1$
- Mixture model is then

$$g(x; \Psi) = \sum_{h=1}^{K} \alpha_h c_p(\eta_h) \exp \{ \eta_h \mu_h^T x \}$$
The Q-function has the following form

\[
Q(\Psi; \Psi^{(b-1)}, y) = \sum_{h=1}^{K} \left( \sum_{i=1}^{n} \pi_{ih} \left\{ \log \alpha_h + \log c_d(\eta_h) + \eta_h \mu_h^T x_i \right\} \right.
\]
\[
+ \lambda_h \left( 1 - \mu_h^T \mu_h \right) \left. \right) 
\]

The parameter estimates at the \(b\)th iteration of the M-step are

\[
\hat{\mu}_h^{(b)} = \frac{r_h^{(b)}}{\|r_h^{(b)}\|}, \quad \alpha_h^{(b)} = \frac{\sum_{i=1}^{n} \pi_{ih}^{(b)}}{n}, \quad \text{and} \quad \frac{c'_d(\eta_h^{(b)})}{c_d(\eta_h^{(b)})} = -\frac{\|r_h^{(b)}\|}{\sum_{i=1}^{n} \pi_{ih}^{(b)}},
\]

where \(r_h = \sum_{i=1}^{n} x_i \pi_{ih}\). A common approximation for \(\eta_h\) is \(\frac{p\bar{r}_h - \bar{r}_h^3}{1 - \bar{r}_h^2}\), where \(\bar{r}_h = \|\sum_{i=1}^{n} x_i \pi_{ih}\|/n\).
Adopting notation from Maugis et al. [9], partition variables \( \{1, \ldots, p\} \):

- \( S^C \): Variables **not** relevant for clustering
- \( W \): Noise
- \( U \): Redundant
- \( S \): Variables relevant for clustering
- \( R \): Explain variables in \( U \)
- \( S \setminus R \)
Variable selection for clustering

Modified EM

EM for redundant variable $s(U&R)$

Redundant Variables: common mean direction $\mu_{hG}$ within each mixture component (denoted by set $G$)

Example: $G = \{1, 2\}$ when $\mu_1 = \frac{1}{\sqrt{2}}\langle 1, 1, 0 \rangle$, $\mu_2 = \frac{1}{3}\langle 2, 2, 1 \rangle$

Modified $Q$-function:

$$Q^*_{R}(\theta \mid x_i) = \sum_{i=1}^{n} \sum_{h=1}^{K} \pi_{ih} \left( \ln(\alpha_h) + \ln(c_p(\kappa_h)) \right)$$

$$+ \sum_{i=1}^{n} \sum_{h=1}^{K} \sum_{j\in G} \pi_{ih} \kappa_h \left( \mu_{hG} \sum_{j\in G} x_{i,j} + \sum_{j\notin G} \mu_{hj} x_{i,j} \right)$$

$$+ \xi \left( 1 - \sum_{h=1}^{K} \alpha_h \right) + \sum_{h=1}^{K} \lambda_h \left[ 1 - |G|\mu_{hG}^2 - \sum_{j\notin G} \mu_{hj}^2 \right]$$

Semhar.Michael@sdstate.edu (SDSU)
EM for Redundant Variables \((U&R)\)

M-Step Updates

\[
\begin{align*}
    r_{hj} &= \sum_{i=1}^{n} \pi_{ih} x_{i,j} \\
    \alpha_h &= \frac{1}{n} \sum_{i=1}^{n} \pi_{ih} \\
    \mu_{h} &= \frac{\sum_{j \in G} r_{hj}}{\sqrt{\left( \frac{1}{|G|} \left( \sum_{j \in G} r_{hj} \right)^2 + \sum_{j \notin G} (r_{hj})^2 \right)}} |G| \\
    \mu_{hj} &= \frac{r_{hj}}{\sqrt{\left( \frac{1}{|G|} \left( \sum_{j \in G} r_{hj} \right)^2 + \sum_{j \notin G} (r_{hj})^2 \right)}} \\
    \kappa_h &\text{ estimated by numerical optimization provided in \textquote{optimize} R function}
\end{align*}
\]
EM for noise variables ($W$)

Noise Variables: common mean direction $\mu_j$ over all components (denoted by set $G$)

Example: $G = \{1, 2\}$ when $\mu_1 = \frac{1}{\sqrt{6}}\langle 1, -1, 0, 2 \rangle$, $\mu_2 = \frac{1}{\sqrt{6}}\langle 1, -1, 2, 0 \rangle$

Modified $Q$-function:

$$Q_N^*(\theta \mid x_i) = \sum_{i=1}^{n} \sum_{h=1}^{K} \pi_{ih} \left(\ln(\alpha_h) + \ln(c_p(\kappa_h))\right)$$

$$+ \sum_{h=1}^{K} \sum_{i=1}^{n} \pi_{ih} \kappa_h \left(\sum_{j \in G} \mu_{.j} x_{i,j} + \sum_{j \notin G} \mu_{hj} x_{i,j}\right)$$

$$+ \xi \left(1 - \sum_{h=1}^{K} \alpha_h\right) + \sum_{h=1}^{K} \lambda_h \left(1 - \sum_{j \in G} \mu^2_{.j} - \sum_{j \notin G} \mu^2_{hj}\right)$$
M-Step Updates

\[ r_{hj} = \sum_{i=1}^{n} \pi_{ih} x_{i,j} \]

\[ \alpha_h = \frac{1}{n} \sum_{i=1}^{n} \pi_{ih} \]

\[ \mu_{hj} = \frac{\kappa_h r_{hj}}{2 \lambda_h} \quad (4) \]

\[ \mu.j = \frac{\kappa_h \sum_{h=1}^{K} r_{hj}}{2 \sum_{h=1}^{K} \lambda_h} \quad (5) \]

\[ 1 = \sum_{j \in G} \mu_{j}^2 + \sum_{j \notin G} \mu_{hj}^2 \]

\( \kappa_h \) estimated by numerical optimization provided in ‘optimize‘ R function
Overview

- **Greedy**: Computationally efficient & Remove variables aggressively
- **Backward Stepwise**: Computationally expensive & remove variables conservatively

All methods use BIC to determine variable type

\[
\text{BIC} = \ln(n) \cdot C - 2 \cdot l(\hat{\theta} \mid x)
\]
Greedy method for $U&R$

**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 $BIC_{\mathcal{M}_0}$

**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 $BIC_{\mathcal{M}_{i,j}}$
- If $BIC_{\mathcal{M}_{i,j}} \leq BIC_{\mathcal{M}_0}$ then
  - Append the pair $i, j$ to $L$
- end

**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, $\mathcal{M}_R$, to $X'$ using the Standard-EM procedure to obtain soft clustering of $X$;

**Algorithm 1:** Greedy vMFM for redundant variables

- **Step 1:** Begin with $S = \{1, \ldots, p\}$ and empty list, $L$ of redundant pairs. Fit $\mathcal{M}_0$ using Standard-EM
- **Step 2:** For each pair of variables, $i, j \in S$, fit $\mathcal{M}_{i,j}$ using Redundant-EM with $G = \{i, j\}$.
  If $BIC_{\mathcal{M}_{i,j}} < BIC_{\mathcal{M}_0}$, Append $i, j$ to $L$
- **Step 3:** Partition $L$ to maximize size of $U$ and minimize size of $R$.
  Fit $\mathcal{M}_R$ using new $S$ and Standard-EM
### Greedy method for $W$

**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 $\text{BIC}_{\mathcal{M}_l} \leq \text{BIC}_{\mathcal{M}_0}$ 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, $\mathcal{M}_R$, to $X'$ using the Standard-EM procedure to obtain soft clustering of $X$;

**Algorithm 2:** Greedy vMFM for noise variables
Simulation setup

- A two-component \((K = 2)\) mixture with ten dimensions \((p = 10)\)
- Simulation Study: Setup
  - 1,000 datasets simulated from \(f(x \mid \theta) = \sum_{h=1}^{2} \alpha_h c_{p}(\kappa_h) e^{\kappa_h \mu_h^T x}\)
  - 1,000 observations from each dataset
- Greedy vMFM Redundant + Greedy vMFM Noise applied
  - \(W = (4, 5)\)
  - \(U = (2, 3)\)
  - \(R = (1)\)
  - \(S = (1, 6, 7, 8)\)

<table>
<thead>
<tr>
<th>(h)</th>
<th>(\alpha_h)</th>
<th>(\kappa_h)</th>
<th>(\mu_h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
<td>4</td>
<td>((-0.49, -0.49, -0.49, -0.16, 0.16, 0, 0, 0.49))</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>4</td>
<td>((0.35, 0.35, 0.35, -0.16, 0.16, 0.47, -0.47, -0.35))</td>
</tr>
</tbody>
</table>
## Results

### Redundant variables

<table>
<thead>
<tr>
<th>Model</th>
<th>BIC</th>
<th>RelDiff(%)</th>
<th>Likelihood</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}_0$</td>
<td>-582.2016</td>
<td>0.0000</td>
<td>356.7245</td>
<td>19</td>
</tr>
<tr>
<td>$\mathcal{M}_{1,2}$</td>
<td>-595.3008</td>
<td>2.2499</td>
<td>356.3663</td>
<td>17</td>
</tr>
<tr>
<td>$\mathcal{M}_{1,3}$</td>
<td>-595.3840</td>
<td>2.2642</td>
<td>356.4079</td>
<td>17</td>
</tr>
<tr>
<td>$\mathcal{M}_{2,3}$</td>
<td>-593.3244</td>
<td>1.9105</td>
<td>355.3781</td>
<td>17</td>
</tr>
<tr>
<td>$\mathcal{M}_{6,7}$</td>
<td>-257.0278</td>
<td>-55.8524</td>
<td>187.2298</td>
<td>17</td>
</tr>
</tbody>
</table>

### Noise Variables

<table>
<thead>
<tr>
<th>Model</th>
<th>BIC</th>
<th>RelDiff(%)</th>
<th>Likelihood</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}_R$</td>
<td>-376.0653</td>
<td>0.0000</td>
<td>239.8408</td>
<td>15</td>
</tr>
<tr>
<td>$\mathcal{M}_4$</td>
<td>-380.9353</td>
<td>1.2950</td>
<td>238.8219</td>
<td>14</td>
</tr>
<tr>
<td>$\mathcal{M}_5$</td>
<td>-377.7816</td>
<td>0.4564</td>
<td>237.2451</td>
<td>14</td>
</tr>
<tr>
<td>$\mathcal{M}_8$</td>
<td>-276.2742</td>
<td>-26.5356</td>
<td>186.4914</td>
<td>14</td>
</tr>
</tbody>
</table>

**Summary redundant**

* 2+ pairs identified in 1000 simulations
* 3 pairs identified in 999 simulations
* No false positives

**Summary for noise**

* 1+ identified in 999 simulations
* 2 identified in 979 simulations
* No false positives
Wisconsin Breast Cancer Data

Description
- 683 observations of clumps of breast cancer cells
- 9 discrete measures
- Benign (444) vs. Malignant (239)

Results
- Stepwise Redundant vMFM + Stepwise Noise vMFM
- 6 redundant
- 0 noise
Conclusion

- Successfully identified noise and Redundant variables in simulation study
- Maintained or improved accuracy in applications

Future development:
- Relax relationship between redundant Variables
- More methods for preselecting likely irrelevant variables
- Computational concerns for identifying noise variables in large datasets
Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S.
Clustering on the unit hypersphere using von mises-fisher distributions.

Bijral, A. S., Breitenbach, M., and Grudic, G.
Mixture of watson distributions: A generative model for hyperspherical embeddings.

 Dortet-Bernadet, J.-L., and Wicker, N.
Model-based clustering on the unit sphere with an illustration using gene expression profiles.

Golzy, M., Markatou, M., and Shivram, A.
Algorithms for clustering on the sphere: Advances & applications.

Guo, J., Levina, E., Michailidis, G., and Zhu, J.
Pairwise variable selection for high-dimensional model-based clustering.
LAW, M. H. C., FIGUEIREDO, M. A. T., AND JAIN, A. K.
Simultaneous feature selection and clustering using mixture models.

LESKOVEC, J., RAJARAMAN, A., AND ULLMAN, J. D.
*Mining of Massive Datasets*, second ed.

MAITRA, R., AND RAMLER, I. P.
A k-mean-directions algorithm for fast clustering of data on the sphere.

MAUGIS, C., CELEUX, G., AND MARTIN-MANGINETTE, M. L.
Variable selection in model-based clustering: A general variable role modeling.

PAN, W., AND SHEN, X.
Penalized model-based clustering with application to variable selection.

RAFTERY, A. E., AND DEAN, N.
Variable selection for model-based clustering.
WANG, S., AND ZHU, J.
Variable selection for model-based high-dimensional clustering and its application to microarray data.
*Biometrics* 64, 2 (2008), 440–448.

ZHONG, S.
Efficient online spherical k-means clustering.