Phase 2 - Comparative Cluster Analysis

Jessala A. Grijalva

Setup

Load Data

```
Dataset dimensions: 4785 rows, 28 columns
```

Build Clustering Matrix

```
Rows used (complete on VARS5): 4785
```

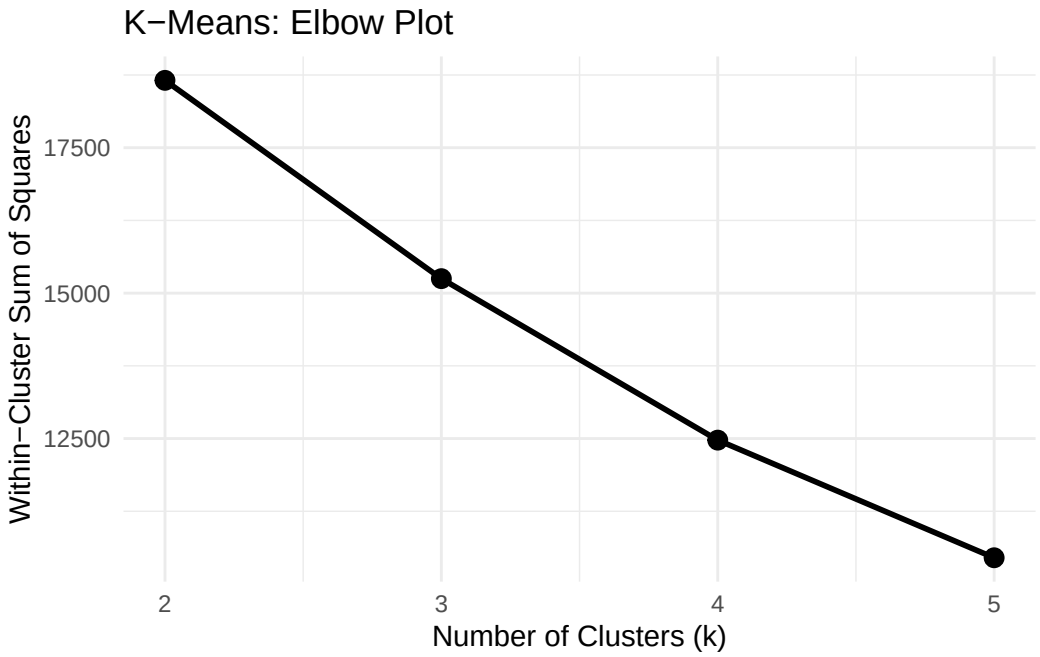
```
Variables: AMERICAN, CULTURAL_IDENTITY, KEEPSPAN, DISTINCT, LEARNENG
```

Step 1: K-Means Clustering

Table 1: K-Means Performance Metrics

k	WCSS	Silhouette	Calinski_Harabasz
2	18657.89	0.3963	1348.96
3	15248.82	0.4040	1359.63
4	12472.70	0.3894	1462.65
5	10450.40	0.4203	1540.24

Elbow Plot



K-Means Cluster Sizes

```
k = 2 : 3723 1062
k = 3 : 997 437 3351
k = 4 : 670 415 2897 803
k = 5 : 405 222 667 2699 792
```

Step 2: Gaussian Mixture Model (GMM)

Table 2: GMM (EEV) Performance Metrics

k	BIC	LogLik	Silhouette	Calinski_Harabasz
2	-58224.86	-28959.91	0.2615	921.54
3	-53579.79	-26569.59	0.3620	1163.86
4	-51688.57	-25556.19	0.3629	1063.27
5	-42298.06	-20793.16	0.3319	656.21

GMM Cluster Sizes

```
k = 2 : 3113 1672
k = 3 : 439 983 3363
k = 4 : 433 705 378 3269
k = 5 : 92 601 175 787 3130
```

GMM k=4 Detail

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EEV (ellipsoidal, equal volume and shape) model with 4 components:

 log-likelihood    n df       BIC       ICL
      -25556.19 4785 68 -51688.57 -52287.66

Clustering table:
   1    2    3    4
 433  705  378 3269
```

Step 3: Fuzzy C-Means

Table 3: Fuzzy C-Means Performance Metrics

k	Silhouette	Calinski_Harabasz	FPC
2	0.3538	1232.47	0.6380
3	0.2816	1214.75	0.5474
4	0.3078	1084.20	0.5038
5	0.3519	1362.46	0.4950

Fuzzy C-Means Cluster Sizes

```
k = 2 : 1538 3247
k = 3 : 1145 1987 1653
k = 4 : 910 2029 928 918
k = 5 : 1960 659 431 904 831
```

Step 4: Hierarchical Clustering

Table 4: Hierarchical Clustering Performance Metrics

k	Silhouette	Calinski_Harabasz
2	0.3559	1112.41
3	0.3930	1280.91
4	0.3814	1353.55
5	0.4057	1425.87

Hierarchical Cluster Sizes

```
k = 2 : 1451 3334
k = 3 : 999 3334 452
k = 4 : 999 2748 586 452
k = 5 : 821 2748 586 452 178
```

Step 5: DBSCAN

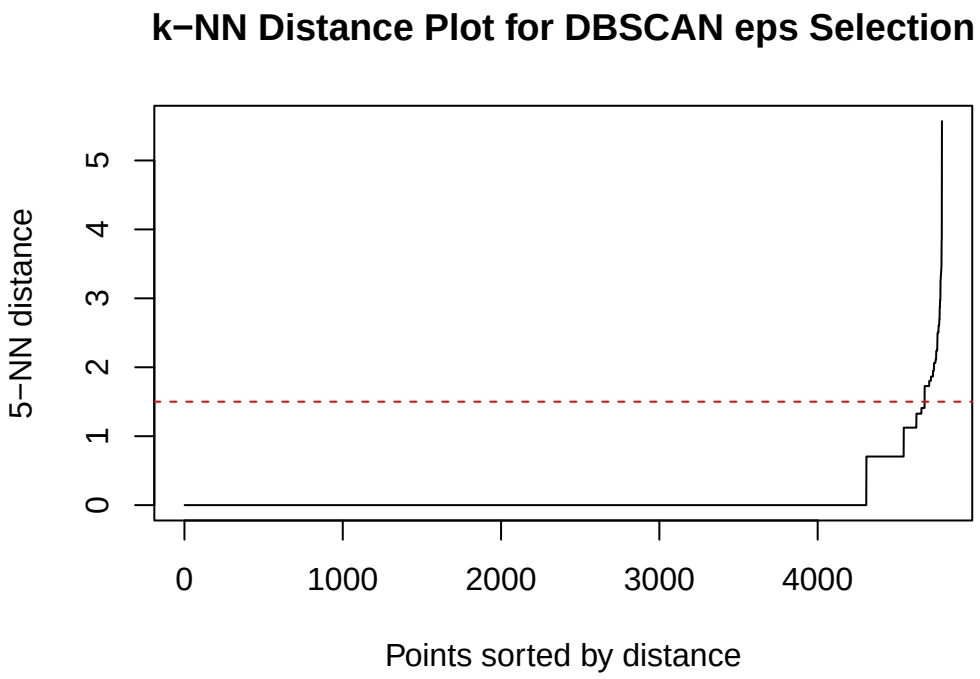


Table 5: DBSCAN Results by eps (minPts = 5)

eps	Clusters	Noise	Pct_Noise	Silhouette
1.00	43	165	3.4	0.4533
1.25	19	113	2.4	0.3831
1.50	21	63	1.3	0.3802
1.75	9	45	0.9	0.3727
2.00	4	25	0.5	0.3578

Comparative Performance Tables

Silhouette Scores by Method and k

Table 6: Silhouette Scores by Method and k

Method	k=2	k=3	k=4	k=5
K-Means	0.3963	0.4040	0.3894	0.4203
GMM (EEV)	0.2615	0.3620	0.3629	0.3319
Fuzzy C-Means	0.3538	0.2816	0.3078	0.3519
Hierarchical	0.3559	0.3930	0.3814	0.4057

Calinski-Harabasz Index by Method and k

Table 7: Calinski-Harabasz Index by Method and k (higher = better)

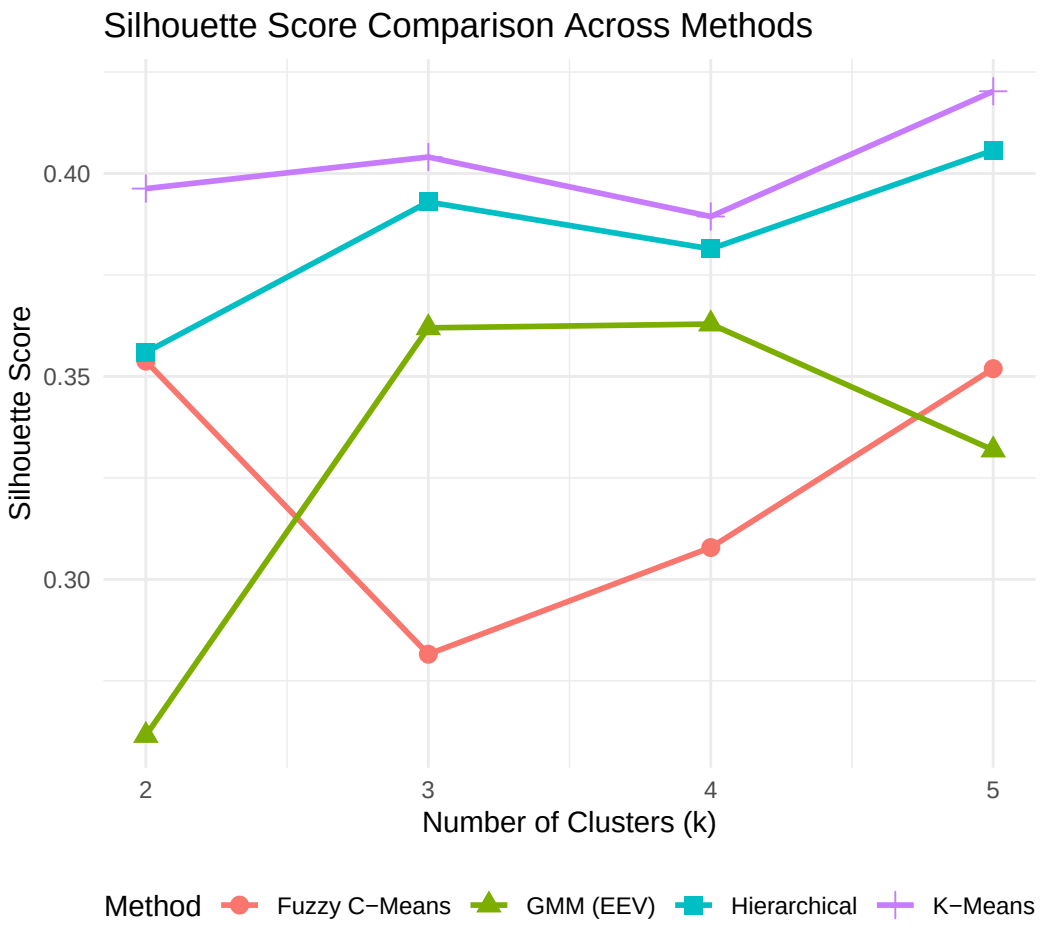
Method	k=2	k=3	k=4	k=5
K-Means	1348.96	1359.63	1462.65	1540.24
GMM (EEV)	921.54	1163.86	1063.27	656.21
Fuzzy C-Means	1232.47	1214.75	1084.20	1362.46
Hierarchical	1112.41	1280.91	1353.55	1425.87

k=4 Solution: Cluster Sizes

Table 8: k=4 Solution: Performance and Cluster Sizes

Method	Silhouette	CH_Index	C1	C2	C3	C4
K-Means	0.3894	1462.65	670	415	2897	803
GMM (EEV)	0.3629	1063.27	433	705	378	3269
Fuzzy C-Means	0.3078	1084.20	910	2029	928	918
Hierarchical	0.3814	1353.55	999	2748	586	452

Performance Visualization



Detailed Comparison: K-Means, GMM, Hierarchical (k=4)

Performance Summary

Table 9: K-Means vs GMM vs Hierarchical at k=4

Method	Silhouette	Calinski_Harabasz	C1	C2	C3	C4
K-Means	0.3894	1462.65	670	415	2897	803
GMM (EEV)	0.3629	1063.27	433	705	378	3269
Hierarchical	0.3814	1353.55	999	2748	586	452

Cluster Means (Standardized Variables)

Table 10: K-Means Cluster Means (k=4)

Cluster	AMERICAN	CULTURAL_IDENTITY	KEEPSPAN	DISTINCT	LEARNENG
1	-1.964	0.046	0.309	0.282	0.224
2	-0.026	-0.006	-0.257	0.028	-2.855
3	0.364	0.269	0.294	0.344	0.295
4	0.338	-1.006	-1.185	-1.492	0.226

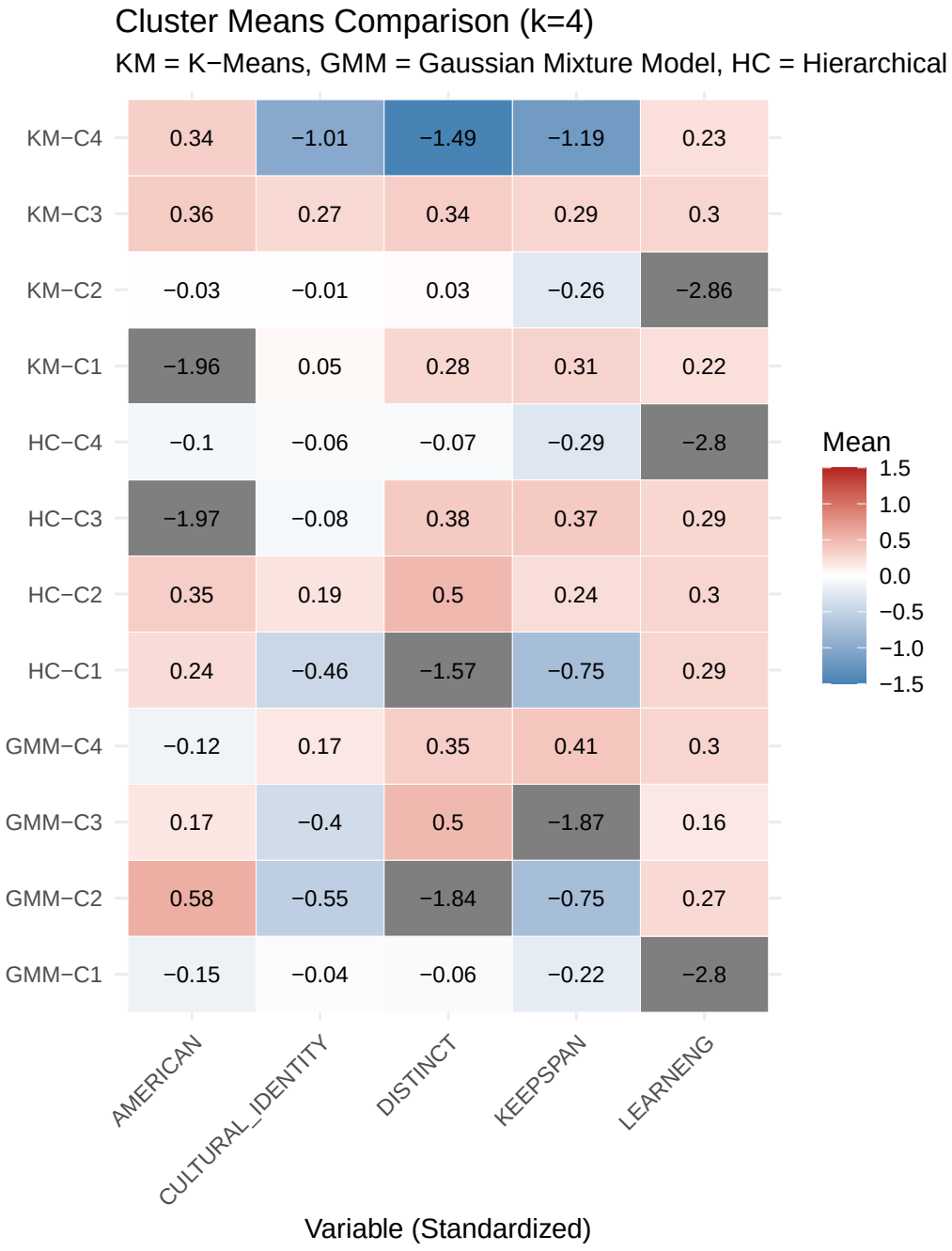
Table 11: GMM (EEV) Cluster Means (k=4)

Cluster	AMERICAN	CULTURAL_IDENTITY	KEEPSPAN	DISTINCT	LEARNENG
1	-0.151	-0.043	-0.218	-0.060	-2.799
2	0.577	-0.548	-0.749	-1.841	0.266
3	0.175	-0.398	-1.872	0.497	0.162
4	-0.125	0.170	0.407	0.348	0.295

Table 12: Hierarchical Cluster Means (k=4)

Cluster	AMERICAN	CULTURAL_IDENTITY	KEEPSPAN	DISTINCT	LEARNENG
1	0.245	-0.455	-0.751	-1.571	0.290
2	0.348	0.192	0.240	0.503	0.295
3	-1.967	-0.083	0.374	0.376	0.286
4	-0.105	-0.056	-0.287	-0.071	-2.803

Cluster Means Heatmap



Summary

```
=== COMPARATIVE CLUSTER ANALYSIS SUMMARY ===

1. GMM (EEV) k=4 REPLICATION:

   Cluster sizes: 433 705 378 3269

   Expected:      433 705 378 3269

2. SILHOUETTE RANKING at k=4:

     K-Means Hierarchical    GMM (EEV)
   0.3893978    0.3814258    0.3629006


3. CALINSKI-HARABASZ RANKING at k=4:

     K-Means Hierarchical    GMM (EEV)
    1462.649     1353.546     1063.266


4. DBSCAN FINDING:

   Does not converge to k=4. Rules out density-based structure.


5. FUZZY C-MEANS FINDING:

   Produces near-equal cluster sizes. Does not match other solutions.

   Low FPC suggests boundaries are not well-defined for fuzzy approach.
```