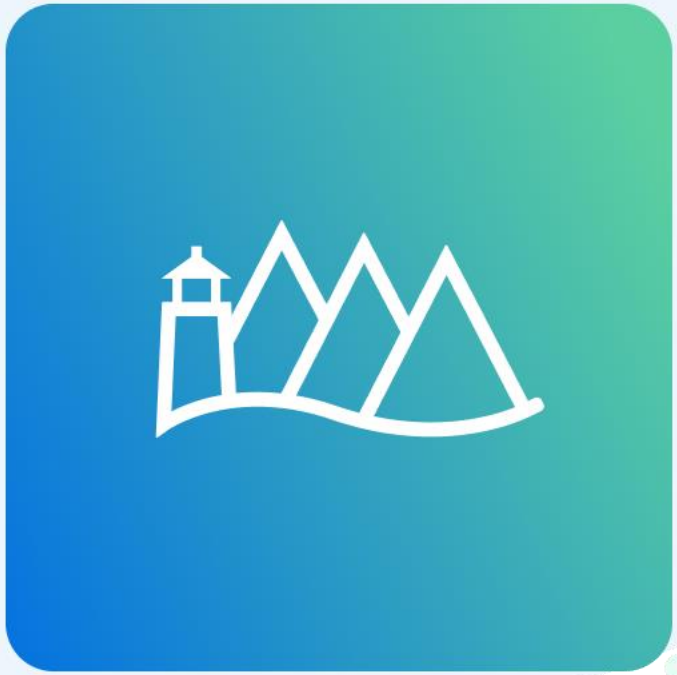
When Segments Misbehave

Segmenting Real-World Marketing Data



Sawtooth Software

Keith Chrzan

Motivation

- Real world segmentation data can be messy in ways that challenge our ability to find the right segments in our data sets
- This webinar focuses on concrete advice to help you meet the challenge

Outline

- Comparing segmentation methods under adverse data conditions
- Comparing variable selection methods under adverse data conditions
- Winners and losers
- Conclusions

Conditions for success

- Clustering methods work best when
  - Segments are
    - Few in number
    - Spherical
    - Of about the same size
  - Your data contains no noise variables, only signal

Which work better under adverse conditions?

- Distance-based methods
  - K-means
    - A robust version using 1,000 random sets of starting seeds
    - A robust version using 50 sets of smart starting seeds
  - Hierarchical
    - Ward's (Agglomerative)
    - Howard-Harris (Divisive)
- Cluster ensemble analysis (using 50 K-means runs as ensemble members)
- Model-based clustering (a finite mixture modeling method, like latent class analysis, appropriate for metric variables)

Artificial data study

- 4,800 data sets
  - Each with 1,000 respondents
  - Each with 5 basis variables
  - Each with segments well-separated in the space (with as few as 2% of cases appearing in overlapping segments in the best conditions and up to 25% in the most adverse conditions)
- 96 experimental cells (50 data sets/condition)
  - 2 segment shape conditions x
  - 2 numbers of segments conditions x
  - 3 segment size conditions x
  - 2 noise conditions x
  - 2 distributional conditions x
  - 2 scale coarsening conditions

Artificial data study – experimental conditions

- Half the data sets have only spherical segments, half have only elliptical
- Half the data sets have 3 segments, half have 6
- One third each of the data sets have equal, moderately imbalanced and severely imbalanced segment sizes
  - For “moderate,” the ratio of the largest sample to the smallest is 5:1
  - It is 10:1 in the “severe condition
- Half of the data sets contain 10 randomly distributed noise variables and half do not
- Half of the data sets are generated from multivariate normal distributions, and half from multivariate skewed and fat-tailed t distributions
- Half of the data sets contain continuous variables, and half contain five-point discretized versions of those same variables

Evaluation criterion: ARI examples

ARI = 1.0
Perfect Agreement

	C1	C2	C3	C4
C1	125	0	0	0
C2	0	125	0	0
C3	0	0	125	0
C4	0	0	0	125

ARI = 0.7
Strong Agreement

	C1	C2	C3	C4
C1	103	11	5	6
C2	6	114	3	2
C3	4	2	116	3
C4	5	7	6	107

ARI = 0.3
Moderate Agreement

	C1	C2	C3	C4
C1	94	8	13	10
C2	18	79	14	14
C3	15	25	78	7
C4	16	13	17	79

ARI = 0.0
No Agreement (Random)

	C1	C2	C3	C4
C1	24	35	29	37
C2	35	21	38	31
C3	29	27	32	37
C4	34	25	23	43

ARI - ideal conditions

- Ideal conditions in this experiment are 3 equal, spherical segments with no masking/noise variables, a Gaussian distribution and continuous variables

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
Ideal	0.586	0.586	0.460	0.479	0.586	0.639

Results – topline

- Mean ARI across all 4,800 data sets in all 96 experimental conditions

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
Average ARI	0.259	0.259	0.204	0.238	0.270	0.344

Results – segment shape

- Mean ARI by segment shape (across all other conditions)

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
Spherical	0.363	0.363	0.283	0.329	0.383	0.416
Elliptical	0.155	0.155	0.124	0.148	0.157	0.271

Results – segment size imbalance

- Mean ARI by relative segment sizes (across all other conditions)

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
None (Equal)	0.300	0.300	0.215	0.264	0.308	0.354
Moderate	0.278	0.277	0.215	0.267	0.290	0.358
Severe	0.199	0.198	0.181	0.184	0.212	0.319

Results – segment number

- Mean ARI by number of segments (across all other conditions)

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
3	0.323	0.322	0.249	0.293	0.327	0.416
6	0.196	0.195	0.158	0.183	0.212	0.271

Results – noise

- Mean ARI by presence of noise/masking variables (across all other conditions)

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
No noise	0.296	0.296	0.267	0.266	0.300	0.384
Noise	0.222	0.221	0.140	0.210	0.239	0.303

Results –normal vs skewed with fat t

- Mean ARI by distributional assumption

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
Gaussian	0.241	0.240	0.183	0.220	0.251	0.346
Skewed, fat-t	0.278	0.277	0.224	0.257	0.289	0.341

Results – continuous vs discretized variables

- Mean ARI by whether data are continuous or discretized like 5-point rating scales

	K-means	K-means++	Wards	Howard-Harris	Ensemble	Model-based
Continuous	0.276	0.276	0.223	0.257	0.285	0.422
Discrete	0.242	0.241	0.185	0.220	0.255	0.265

Conclusions – segmentation methods

- There are a lot of things that are out of your control
  - Number of segments
  - Shape of segments
  - Relative size of segments
  - Variable distribution
  - Variable scaling
- Two important things ARE under the analyst's control
  - Model-based clustering protects against a variety of data pathologies
  - You can reduce the impact of noise/masking variables by variable selection

Variable selection – candidate methods

- Possible solutions
  - Throw ‘em all in
  - “Tandem clustering” using factor analysis to reduce the number of variables
  - Unsupervised random forests to cull noise variables
  - Automatic variable selection with VarSelLCM

Variable selection study

- 4,800 data sets
  - Each with 1,000 respondents
  - Each with 5 signal variables driving cluster structure – these should be the bases
  - Each with segments well-separated in the space
- 48 experimental cells (100 data sets/condition)
  - 2 segment shape conditions x
  - 2 numbers of segments conditions x
  - 3 segment size conditions x
  - 2 noise conditions x
  - 2 redundant variables conditions

Artificial data study – experimental conditions

- Half the data sets have only spherical segments, half have only elliptical
- Half the data sets have 3 segments, half have 6
- One third each of the data sets have equal, moderately imbalanced and severely imbalanced segment sizes
  - For “moderate,” the ratio of the largest sample to the smallest is 5:1
  - It is 10:1 in the “severe condition
- Half of the data sets contain 10 randomly distributed noise variables and half contain 50
- Half the data sets contain 20 new variables correlated at about 0.65 with randomly selected signal and noise variables and half do not

Methods tested

- Oracle → mclust
  - Enter all and only the 5 true signal variables into mclust model-based (“latent class”) clustering
- All-in → mclust
- FA → mclust
- URF w/Boruta → mclust
- URF-MVN w/ Boruta → mclust
- VarSelLCM → mclust

Methods tested

- Oracle → mclust
- All-in → mclust
  - Use all the variables in mclust (i.e., no variable selection at all)
  - The “garbage can” approach
- FA → mclust
- URF w/Boruta → mclust
- URF-MVN w/ Boruta → mclust
- VarSelLCM → mclust

Methods tested

- Oracle → mclust
- All-in → mclust
- FA → mclust
  - Run “parallel” method to identify the number of factors (Horn 1965)
  - Then run mclust using the factor scores as bases
- URF w/Boruta → mclust
- URF-MVN w/ Boruta → mclust
- VarSelLCM → mclust

Methods tested

- Oracle
- All-in → mclust
- FA → mclust
- URF w/Boruta → mclust
  - Use Boruta (Kursa & Rudnicki 2010) to identify significant predictors in unsupervised random forest (Brieman 2000, Shi & Horvath 2006) analysis
  - Then mclust using only those variables to make segments
- URF-MVN w/ Boruta → mclust
- VarSelLCM → mclust

Methods tested

- Oracle → mclust
- All-in → mclust
- FA → mclust
- URF w/Boruta → mclust
- URF-MVN w/ Boruta → mclust
  - Same as above but with a customized version of URF called URF-MVN
    - Instead of breaking all structure in the fake data portion of URF, we break only segment structure
    - Signal variables identified as those that detect segment structure
- VarSelLCM → mclust

Methods tested

- Oracle → mclust
- All-in → mclust
- FA → mclust
- URF w/Boruta → mclust
- URF-MVN w/ Boruta → mclust
- VarSelLCM → mclust
  - Use VarSelLCM to identify signal variables
  - Then mclust to make segments

Results – ideal

- Ideal is 3 equal-sized spherical segments with no redundant or masking variables

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
Overall	0.637	0.627	0.562	0.350	0.633	0.628

Results – across all cells

- URF-MVN and VarSelLCM have the highest mean ARIs

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
Overall	0.542	0.104	0.166	0.132	0.267	0.281

Results – by segment size balance

- Average ARI across all other conditions . . .

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
Balance: Equal	0.547	0.101	0.174	0.139	0.275	0.294
Balance: Moderate	0.550	0.101	0.172	0.135	0.273	0.288
Balance: Severe	0.530	0.110	0.152	0.123	0.251	0.262

- Surprisingly, while this condition has a strong effect across *segmentation methods* (k-means, ensemble, latent class) it doesn't distinguish as strongly among *variable selection methods*

Results – by number of segments

- Average ARI across all other conditions . . .

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
True K: 3	0.662	0.143	0.223	0.176	0.365	0.358
True K: 6	0.423	0.065	0.109	0.088	0.168	0.205

- Larger numbers of segments harm all the methods, but URF-MVN and VarSelLCM the least

Results – by number of masking variables

- Average ARI across all other conditions . . .

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
Noise: Low (10 vars)	0.543	0.206	0.189	0.188	0.385	0.294
Noise: High (50 vars)	0.542	0.001	0.142	0.077	0.148	0.269

- Masking variables are damaging, and their presence shows why the All-in method is not a viable option for garbage can segmentations

Results – by redundant variables vs not

- Average ARI across all other conditions . . .

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
Redundant: None	0.544	0.181	0.249	0.165	0.384	0.353
Redundant: Present	0.541	0.026	0.082	0.100	0.149	0.210

- Redundant variables are injurious to all methods
- I usually follow the variable selection analysis with correlation analysis to prune redundant variables

Results – by segment shape

- Average ARI across all other conditions . . .

Condition	Oracle	All-in	FA	URF	URF-MVN	VarSelLCM
Shape: Spherical	0.531	0.136	0.240	0.172	0.323	0.393
Shape: Elliptical	0.554	0.072	0.092	0.092	0.211	0.170

- Elliptical segments render all methods other than URF and VarSelLCM almost completely useless

Conclusions – variable selection

- For garbage can clustering with metric variables, use VarSelLCM or maybe URF-MVN with Boruta for variable selection
- This analysis considered garbage can segmentations
  - It does not apply to well thought-out segmentations with small numbers of bases (like 20 or fewer)
    - For fewer than 20 or so bases, maybe do the All-in approach, remove variables that don't differ across segments and then rerun the segmentation again
  - URF-MVN will fail if all the variables are signal and none are noise
    - Use VarSelLCM instead if you're pretty sure you're not facing a garbage can

Ending note: The ontology of segmentation

- What *kind* of thing is a market segment?
  - A real entity that exists in the world – segment realist view
  - A useful fiction we impose on data – segment instrumentalist view
- For the segmentation *realist* – VarSelLCM with model-based clustering gets us closer to the truth, under a wide variety of conditions, than do other variable selection methods or clustering with distance-based measures or ensembles
- Still, there's no reason not to use strong methods, even if you're a segmentation instrumentalist
  - Segmentation instrumentalists might want to check out the article "AI-assisted segmentation" by Guthart *et al.* in the *2025 Sawtooth Research Conference Proceedings* on our website
  - They push the instrumentalist view in a really interesting direction

That’s it

QUESTIONS?

OR, FEEL FREE TO CONTACT ME LATER: KEITH@SAWTOOTH.COM