Variable Selection for Consistent Clustering

Ron Yurko Rebecca Nugent Sam Ventura

Department of Statistics
Carnegie Mellon University

Classification Society, 2017

Typical Questions in Cluster Analysis

How many clusters?

Which variables should we use?

Which clustering method?

Typical Questions in Cluster Analysis

How many clusters?

- Model-Based Cluster Analysis *(Fraley & Raftery, 1998)*
- Gap Statistic *(Tibshirani et al., 2001)*
- Maximum Clustering Similarity
  (Albatineh & Niewiadomska-Bugaj, 2011)

Which variables should we use?

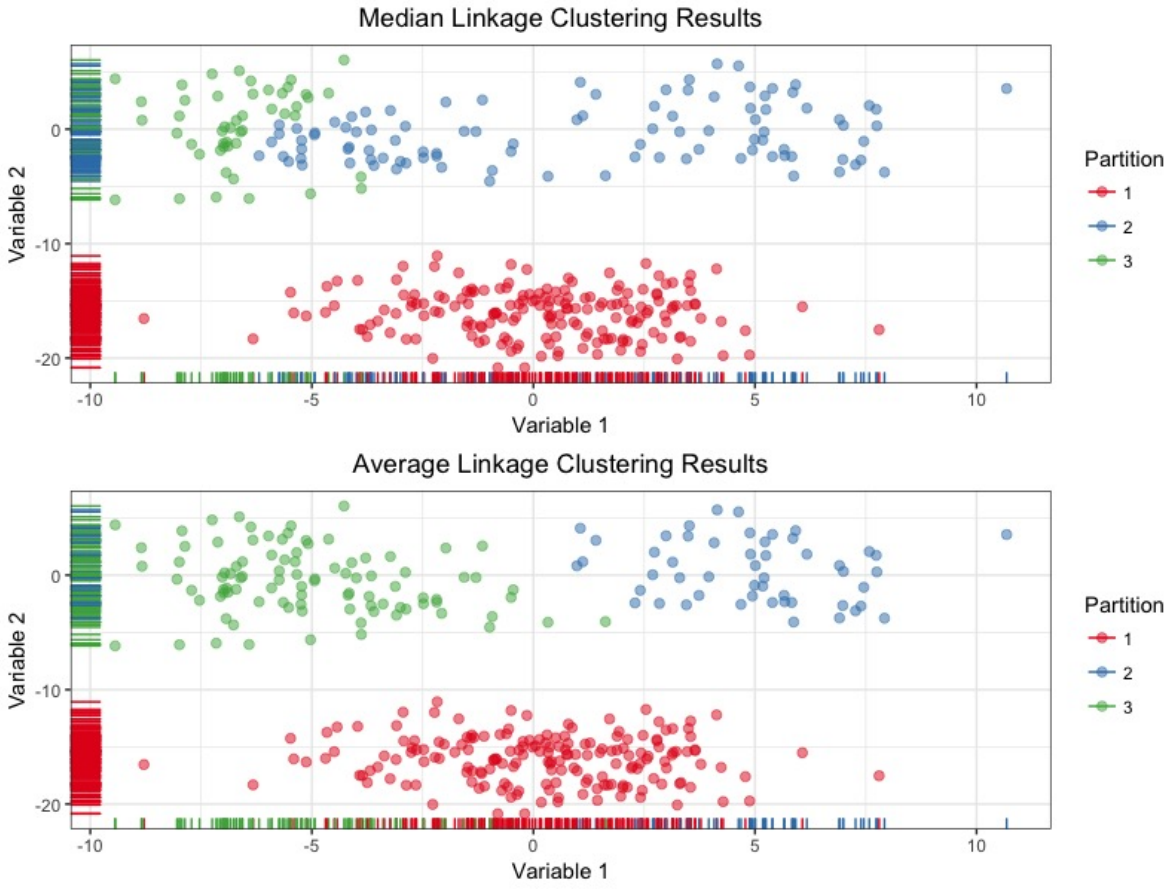
- HINoV Method *(Carmone et al., 1999)*
- Heuristic for K-Means *(Brusco & Cradit, 2001)*
- Model-Based Clustering *(Raftery & Dean, 2006)*

Which clustering method?

- Clustering Validation *(Milligan, 1996)*
- Comparison of Heuristic Procedures *(Brusco & Steinley, 2007)*

Inconsistent Clustering

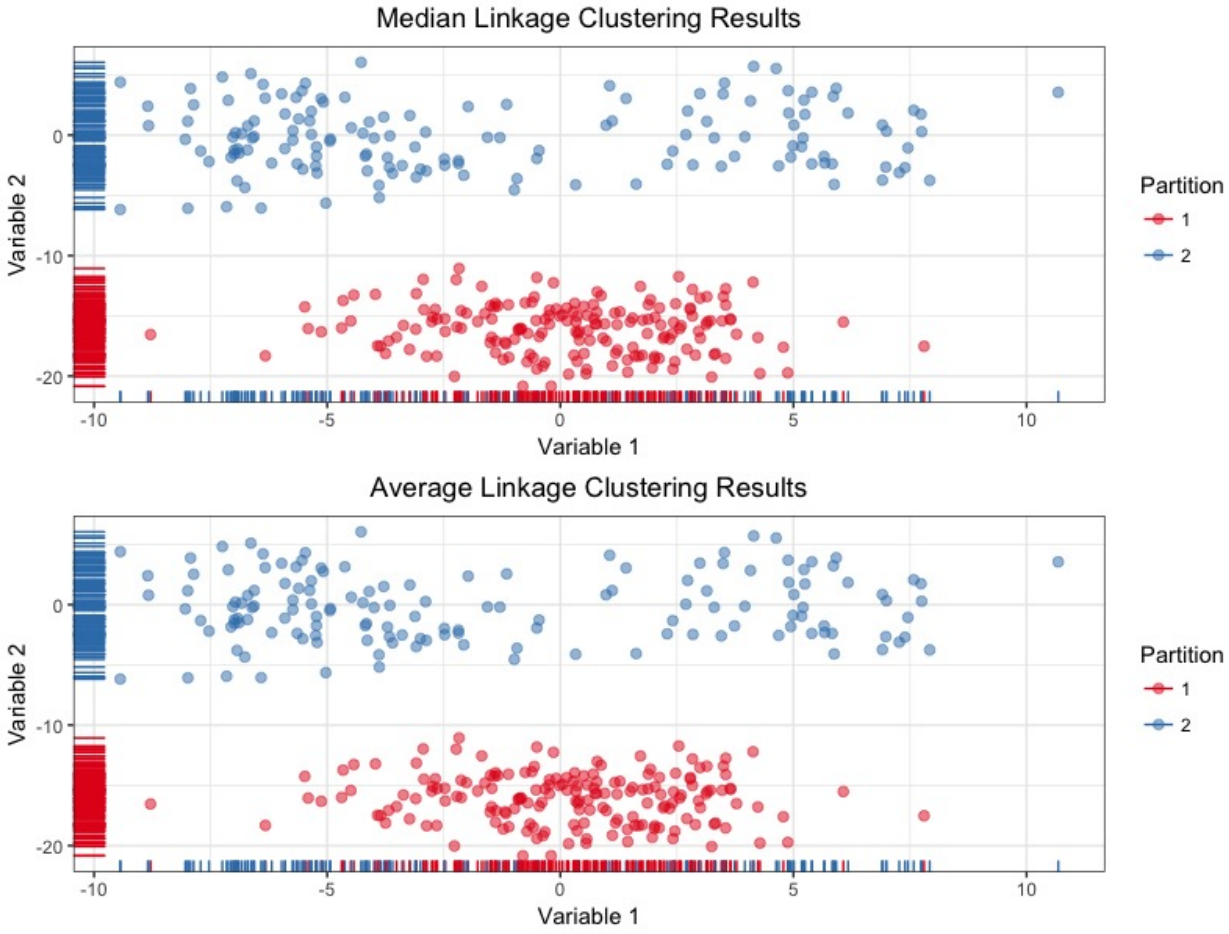
Clustering with both variables and 3 clusters



Disagree on 2 of the 3 clusters,
probably since they're not well separated along Variable 1

Consistent Clustering

Clustering only with Variable 2 and 2 clusters



Perfectly agree on 2 well separated clusters

Variable Selection for Consistent Clustering

GOAL:

Search for the variables yielding consistent clusters based on the level of agreement between methods

Variable Selection for Consistent Clustering

GOAL:

Search for the variables yielding consistent clusters based on the level of agreement between methods

We are NOT optimizing for recovery of "true cluster labels"

We ARE optimizing for agreement of obvious group structure

Variable Selection for Consistent Clustering

Combine variable selection and method agreement

Variable Selection for Consistent Clustering

Combine variable selection and method agreement

Variable Selection:

- Greedy search - find best variable for clustering, given that variable find the next best, repeat
- *Variable Selection for Model-Based Clustering (Raftery & Dean, 2006)*

Variable Selection for Consistent Clustering

Combine variable selection and method agreement

Variable Selection:

- Greedy search - find best variable for clustering, given that variable find the next best, repeat
- *Variable Selection for Model-Based Clustering (Raftery & Dean, 2006)*

Method Agreement:

- Measure agreement between methods using similarity indices to compare clustering partitions
- Adjusted Rand Index (ARI) *(Hubert & Arabie, 1985)*, expected value of 0 and closer to 1 the better
- *MCS: A Method for Finding the Number of Clusters (Albatineh & Niewiadomska-Bugaj, 2011)*

Algorithm Notation

- $\mathbf{X} = N \; x \; D$ data matrix, where $d \in \{1, \ldots, D\}$
- $M =$ set of clustering methods considered, where $m \in \{1, \ldots, |M|\}$
- $S =$ set of variables selected for inclusion
- $U =$ set of unselected variables, where $S \cup U = \{1, \ldots, D\}$ and $S \cap U = \{\emptyset\}$

Algorithm Notation

- $p_{m,S}$ = $(1\ x\ N)$-dimensional row vector that defines a partition with method m given set of variables S
- $ARI(p_{m_i,S},\ p_{m_j,S})$ = ARI between partitions of two different methods m_i and m_j
- $\overline{ARI}_S = \frac{1}{\binom{|M|}{2}} \sum_{i \neq j} ARI(p_{m_i,S},\ p_{m_j,S})$ measures the **consistency** for set of variables S

Algorithm

Step 0: Initialize $S = \{\emptyset\}$, $U = \{1, \ldots, D\}$

Algorithm

Step 0: Initialize $S = \{\emptyset\}$, $U = \{1, \ldots, D\}$

Step 1: For each variable $d \in U$:

- Create partitions with each clustering method
  $p_{m_1, S \cup \{d\}}, \cdots p_{m_{|M|}, S \cup \{d\}}$
- Compute $ARI(p_{m_i, S \cup \{d\}},\ p_{m_j, S \cup \{d\}})$ for each of the $\binom{|M|}{2}$ pairs of partitions

Algorithm

Step 0: Initialize $S = \{\emptyset\}$, $U = \{1, \ldots, D\}$

Step 1: For each variable $d \in U$:

- Create partitions with each clustering method
  $p_{m_1, S \cup \{d\}}, \cdots p_{m_{|M|}, S \cup \{d\}}$
- Compute $ARI(p_{m_i, S \cup \{d\}},\ p_{m_j, S \cup \{d\}})$ for each of the $\binom{|M|}{2}$ pairs of partitions

Step 2: Choose the most consistent variable:

$$d^* := \arg\max_{d \in U} \overline{ARI}_{S \cup \{d\}}$$

Algorithm

Step 0: Initialize $S = \{\emptyset\}$, $U = \{1, \ldots, D\}$

Step 1: For each variable $d \in U$:

- Create partitions with each clustering method $p_{m_1, S \cup \{d\}}, \cdots p_{m_{|M|}, S \cup \{d\}}$
- Compute $ARI(p_{m_i, S \cup \{d\}}, \ p_{m_j, S \cup \{d\}})$ for each of the $\binom{|M|}{2}$ pairs of partitions

Step 2: Choose the most consistent variable:

$$d^* := \arg\max_{d \in U} \overline{ARI}_{S \cup \{d\}}$$

Step 3: Update $S = S \cup \{d^*\}$ and $U = U \backslash \{d^*\}$

Algorithm

Step 0: Initialize $S = \{\emptyset\}$, $U = \{1, \ldots, D\}$

Step 1: For each variable $d \in U$:

- Create partitions with each clustering method $p_{m_1, S\cup\{d\}}, \cdots p_{m_{|M|}, S\cup\{d\}}$
- Compute $ARI(p_{m_i, S\cup\{d\}},\ p_{m_j, S\cup\{d\}})$ for each of the $\binom{|M|}{2}$ pairs of partitions

Step 2: Choose the most consistent variable:

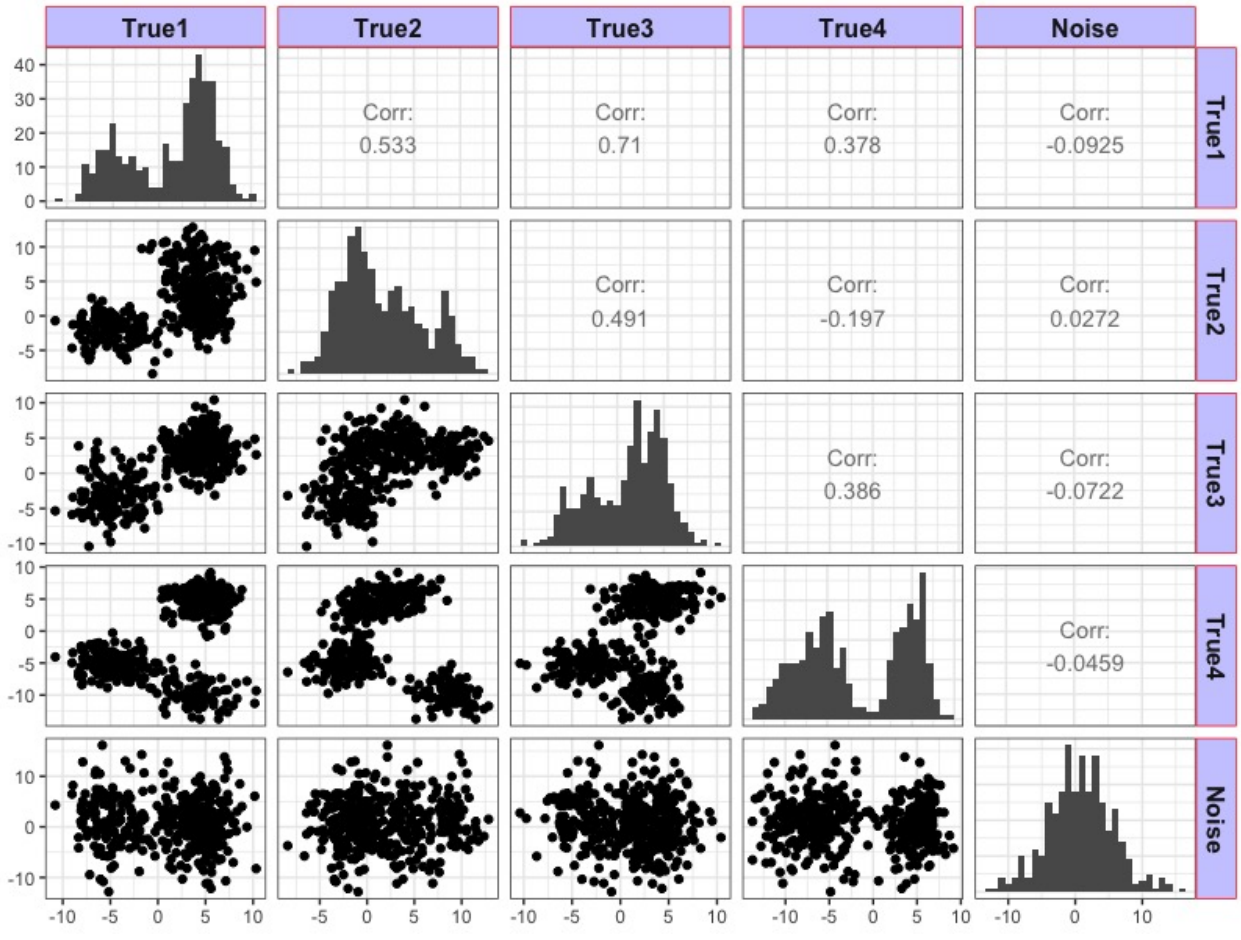
$$d^* := \arg\max_{d\in U} \overline{ARI}_{S\cup\{d\}}$$

Step 3: Update $S = S \cup \{d^*\}$ and $U = U \backslash \{d^*\}$

Repeat Steps 1-3 until including d^* meets stopping criteria

Demo Data

402 observations, 5 variables (4 true, 1 noise),
3 known clusters with moderate separation



Generated with `clusterGeneration` *(Qiu & Joe, 2006)*

Algorithm Demo

For now clustering with known number of 3 clusters

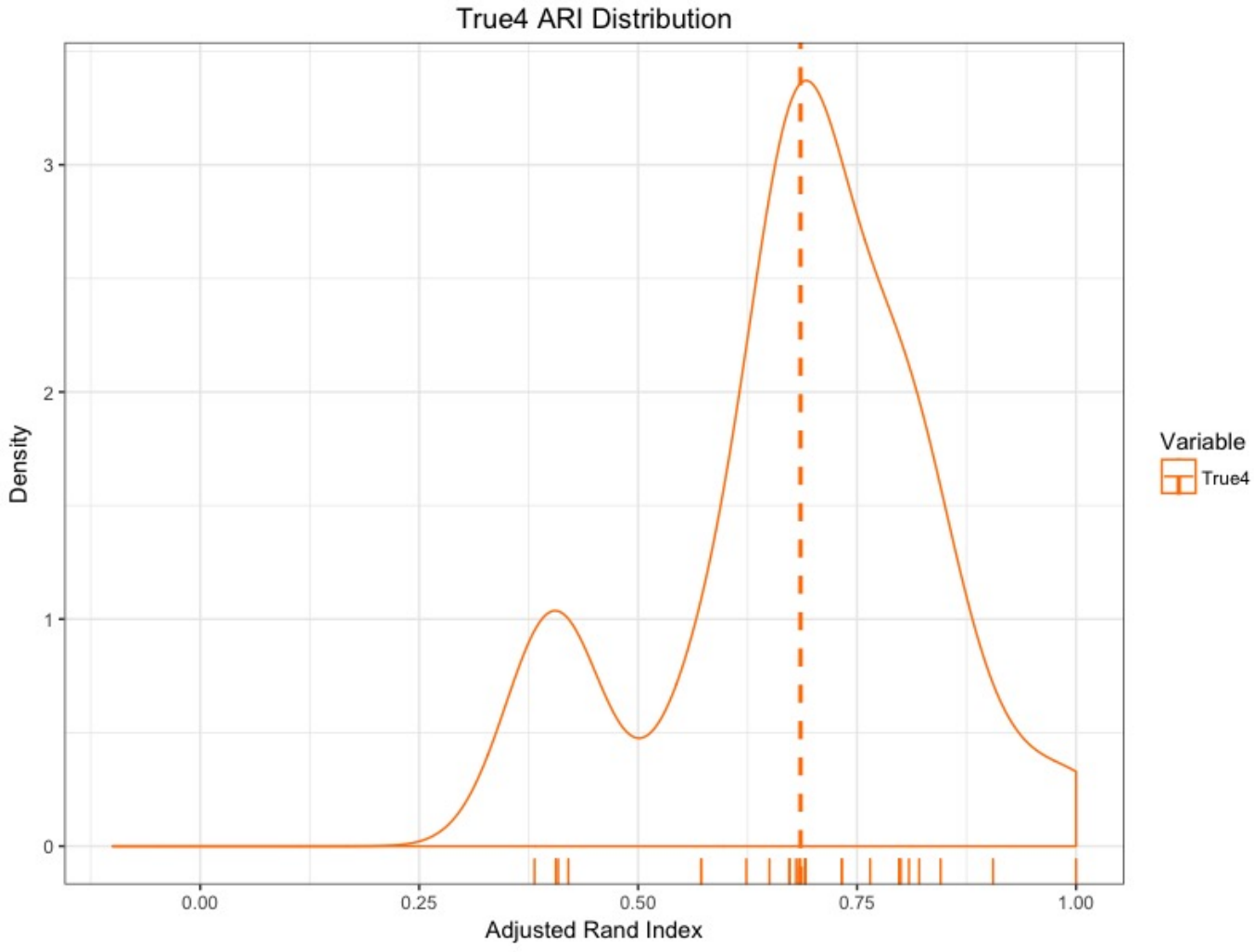
Let $M =$
$\{complete, single, Ward, average, McQuitty, median, centroid, kmeans\}$

For each variable, ARI distribution has $\binom{8}{2} = 28$ values:

Variable	**Method 1**	**Method 2**	**ARI**
True1	Complete	K-Means	0.4583
True2	Complete	K-Means	0.7799
True3	Complete	K-Means	0.3246
True4	Complete	K-Means	0.6803
Noise	Complete	K-Means	0.6400
True1	Complete	Average	0.7195
⋮	⋮	⋮	⋮

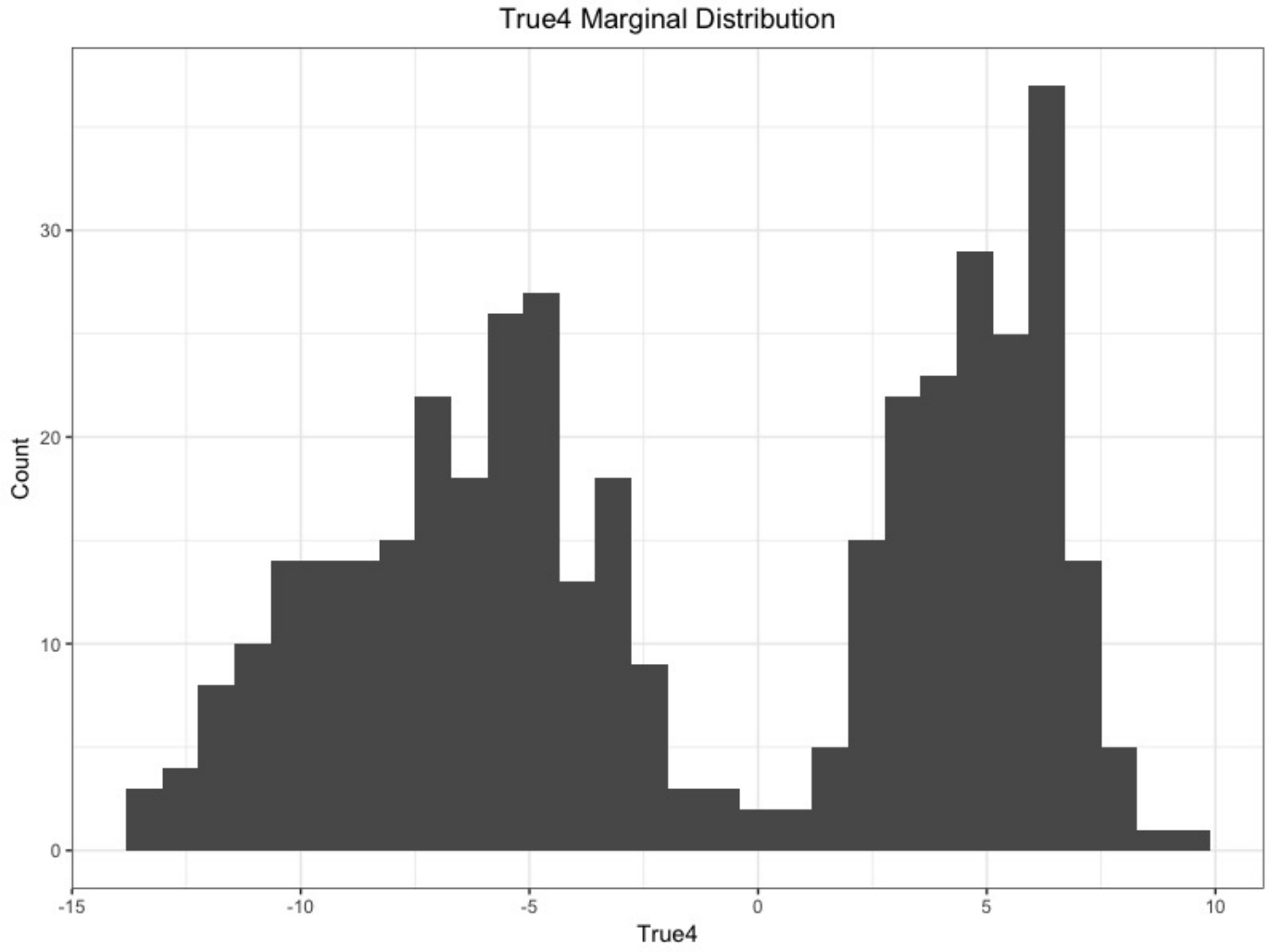
Most Consistent Variable: True4

True4's ARI distribution and dashed line at $\overline{ARI}_{\{True4\}} = 0.6854$



Revisit True4 Marginal Distribution

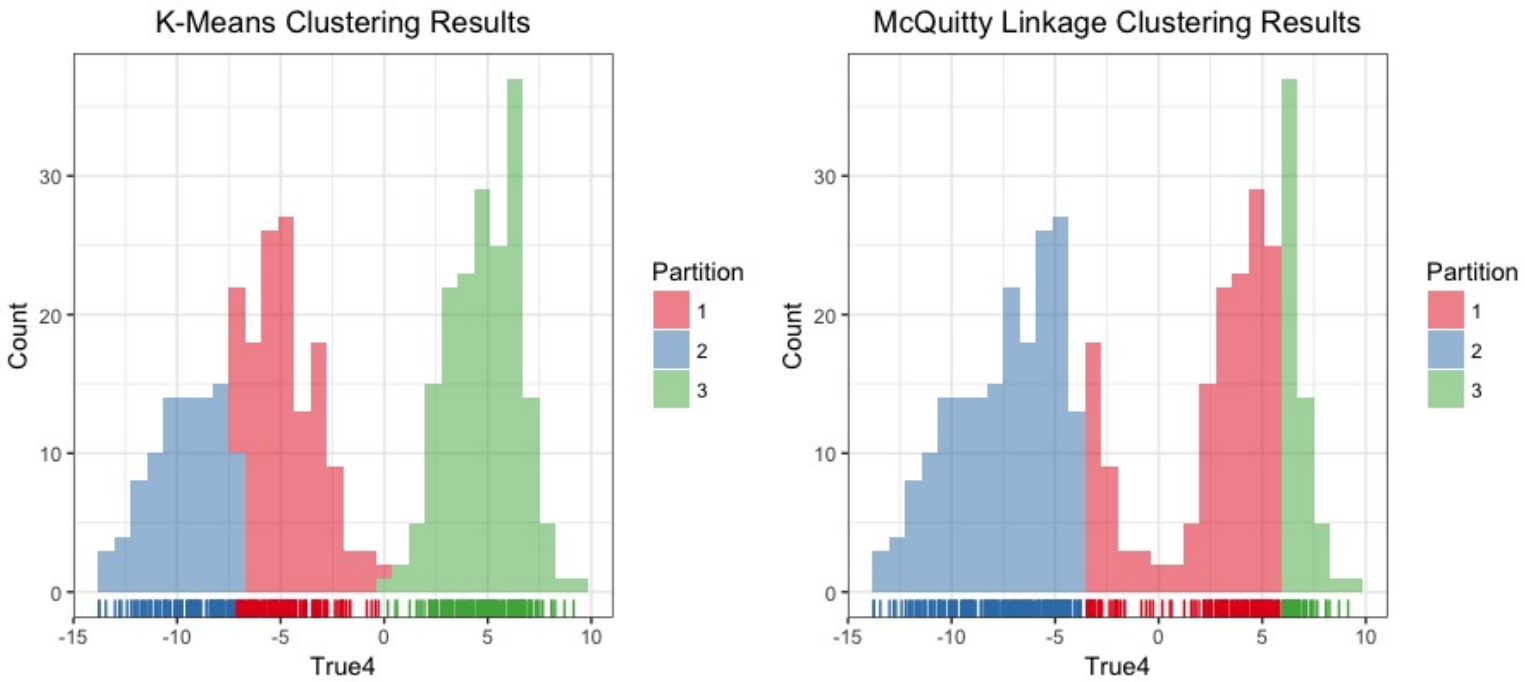
Clear bimodal distribution, but asking for 3 clusters



Clustering Disagreement Example

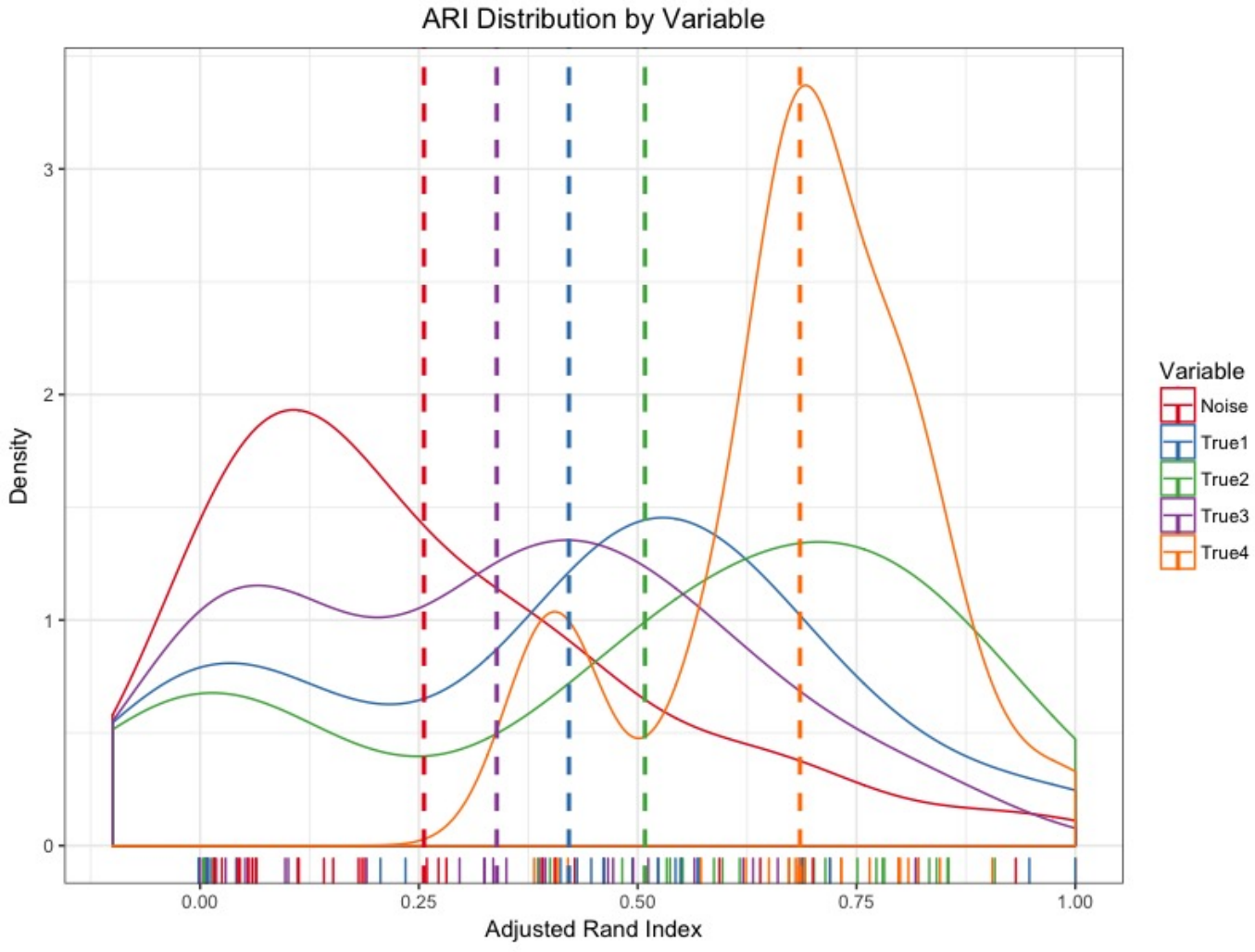
$ARI(p_{kmeans,\{True4\}},\ p_{McQuitty,\{True4\}}) = 0.3816$

Lower mode of ARI values for True4 all involved McQuitty



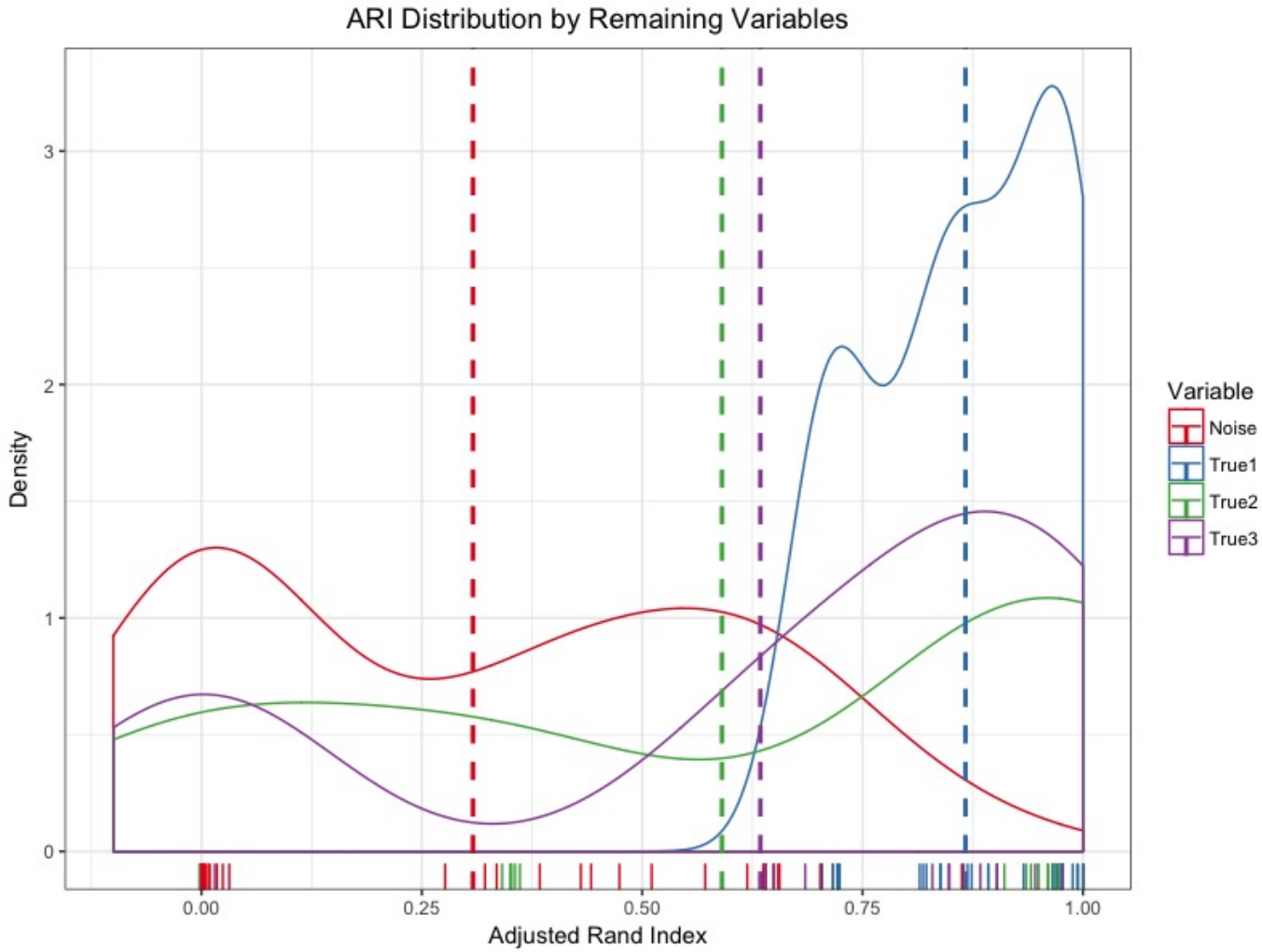
Comparison of ARI Distributions

Can easily view how other variables compare to True4



Given {True4}, Most Consistent: True1

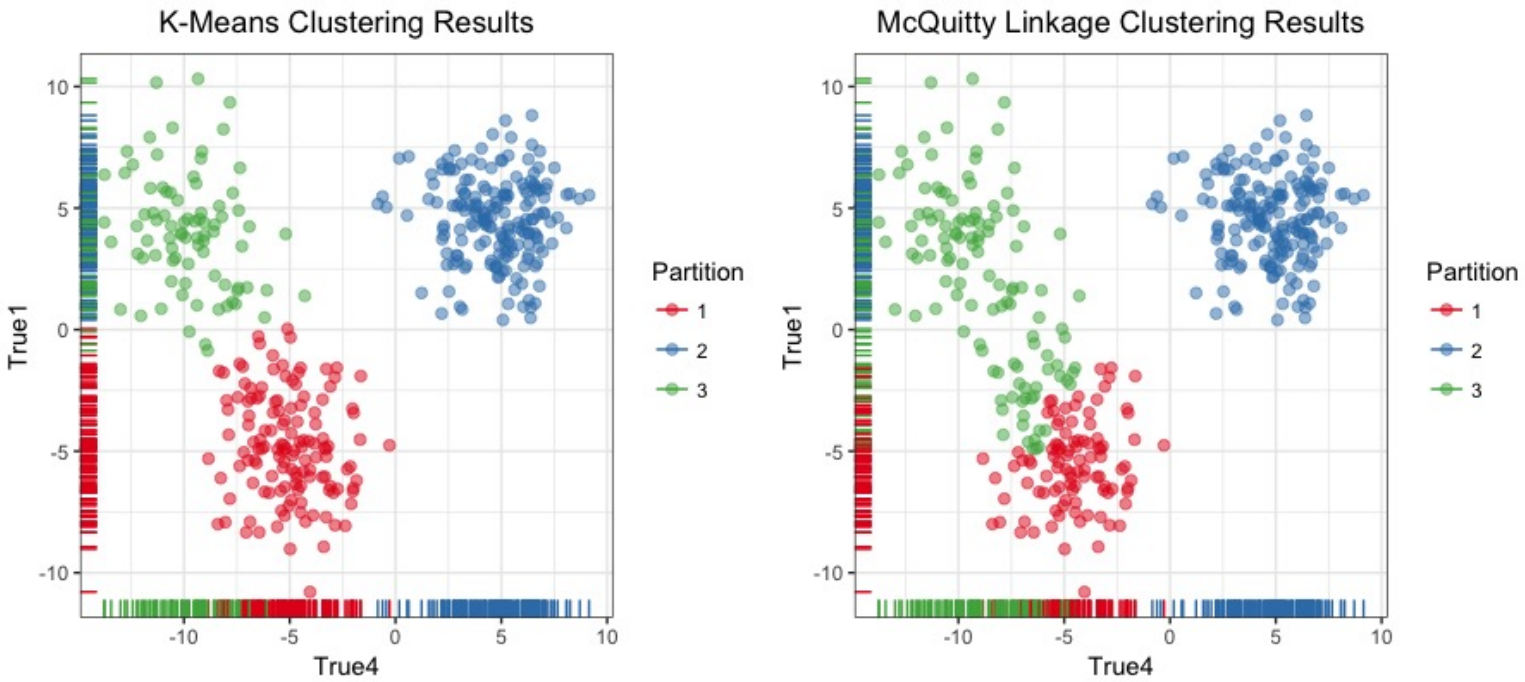
More consistent results, $\overline{ARI}_{\{True4, True1\}} = 0.8668$



Change In Clustering Disagreement

$$ARI(p_{kmeans,\{True4,True1\}},\ p_{McQuitty,\{True4,True1\}}) = 0.8224$$

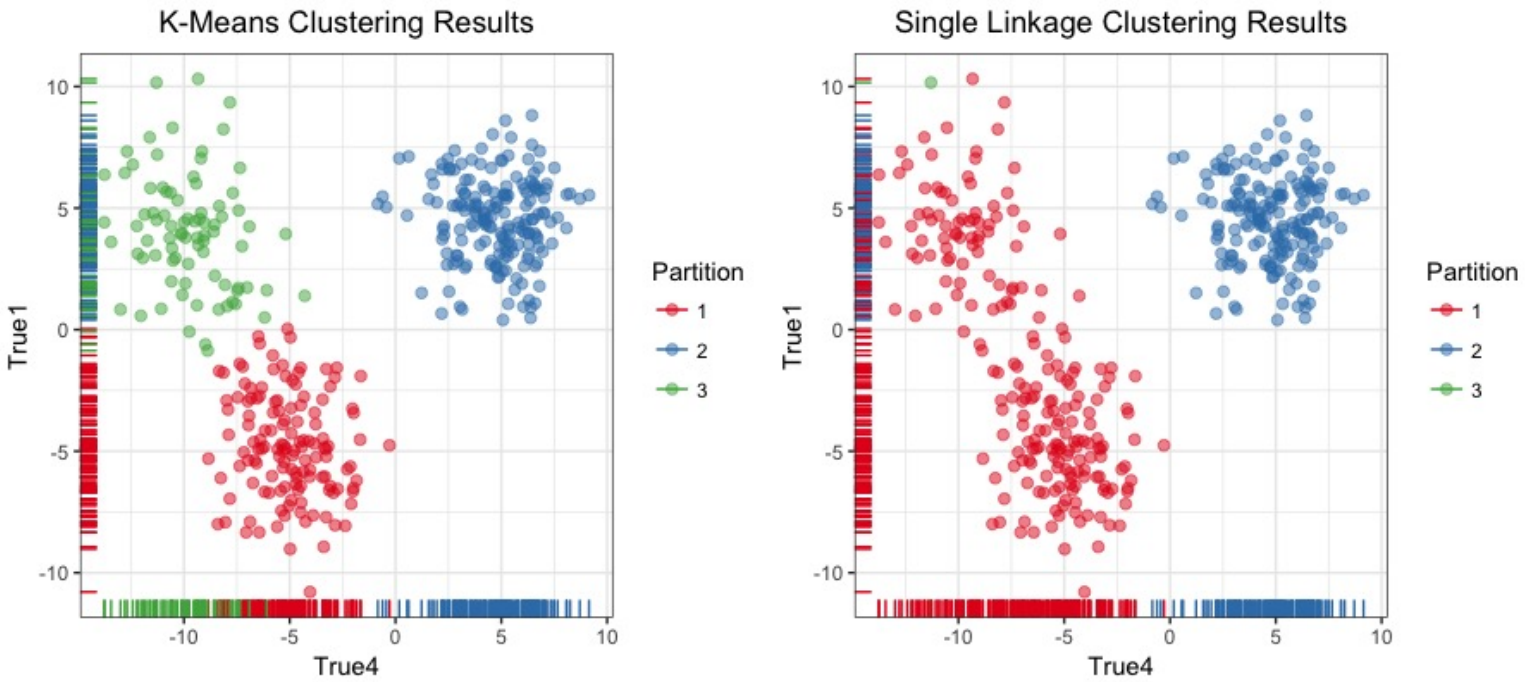
McQuitty is no longer the lower mode of ARI values...



Change In Clustering Disagreement

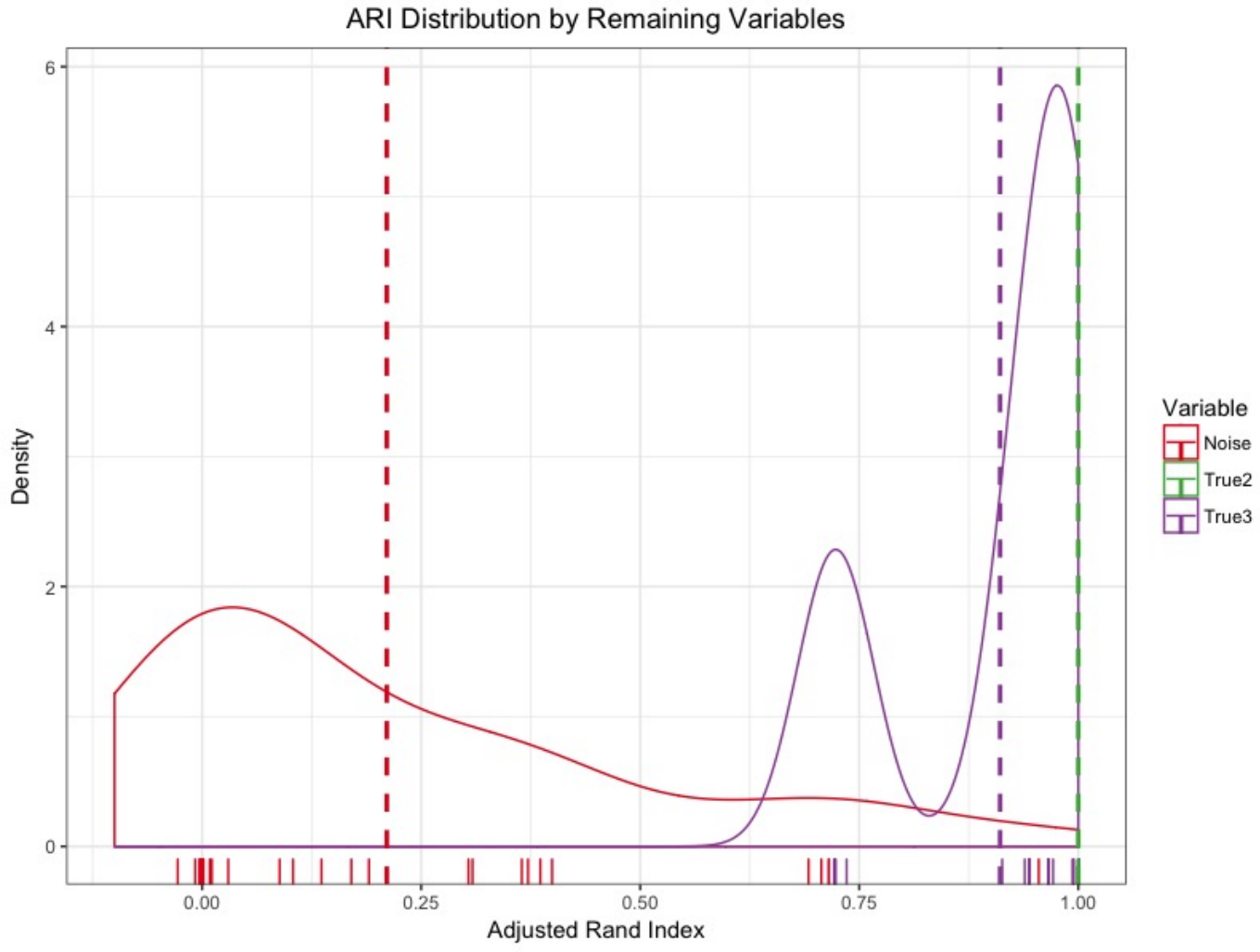
$$ARI(p_{kmeans,\{True4,True1\}},\ p_{single,\{True4,True1\}}) = 0.7213$$

Single linkage makes up the lower mode



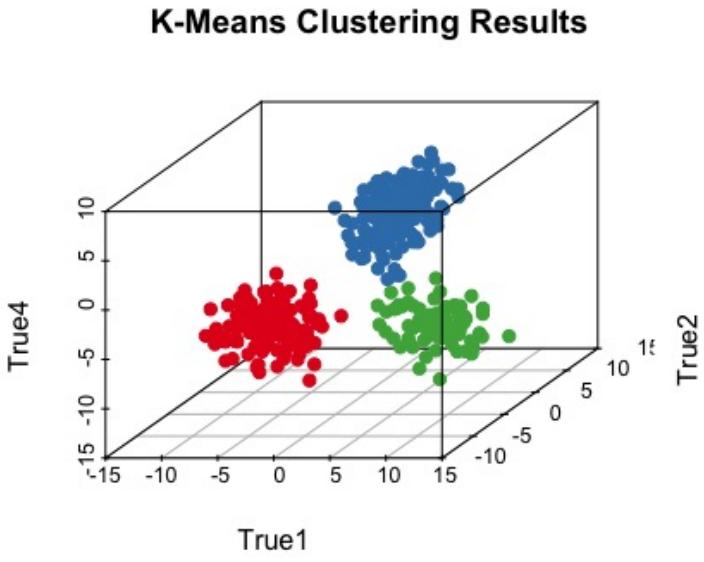
Given {True4, True1}, Most Consistent: True2

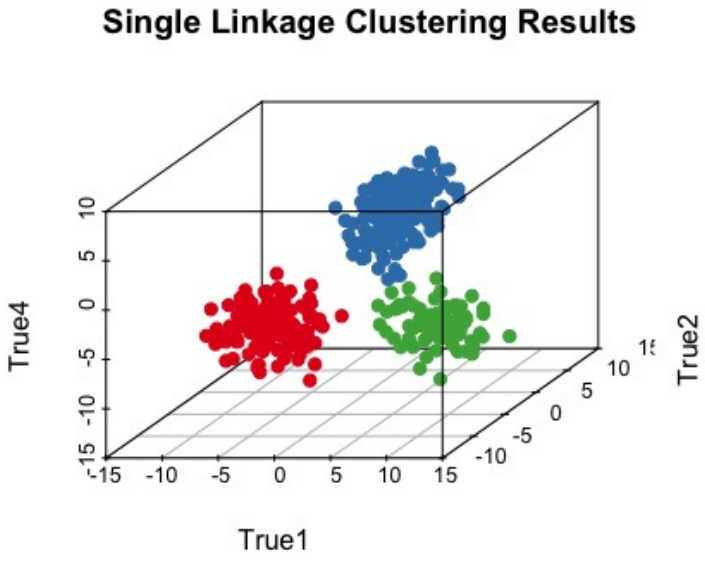
Perfect agreement, $\overline{ARI}_{\{True4, True1, True2\}} = 1$



Consistent Clustering Agreement

$ARI(p_{m_1,\{True4,True1,True2\}},\ p_{m_2,\{True4,True1,True2\}}) = 1,$
across all 28 pairs of methods

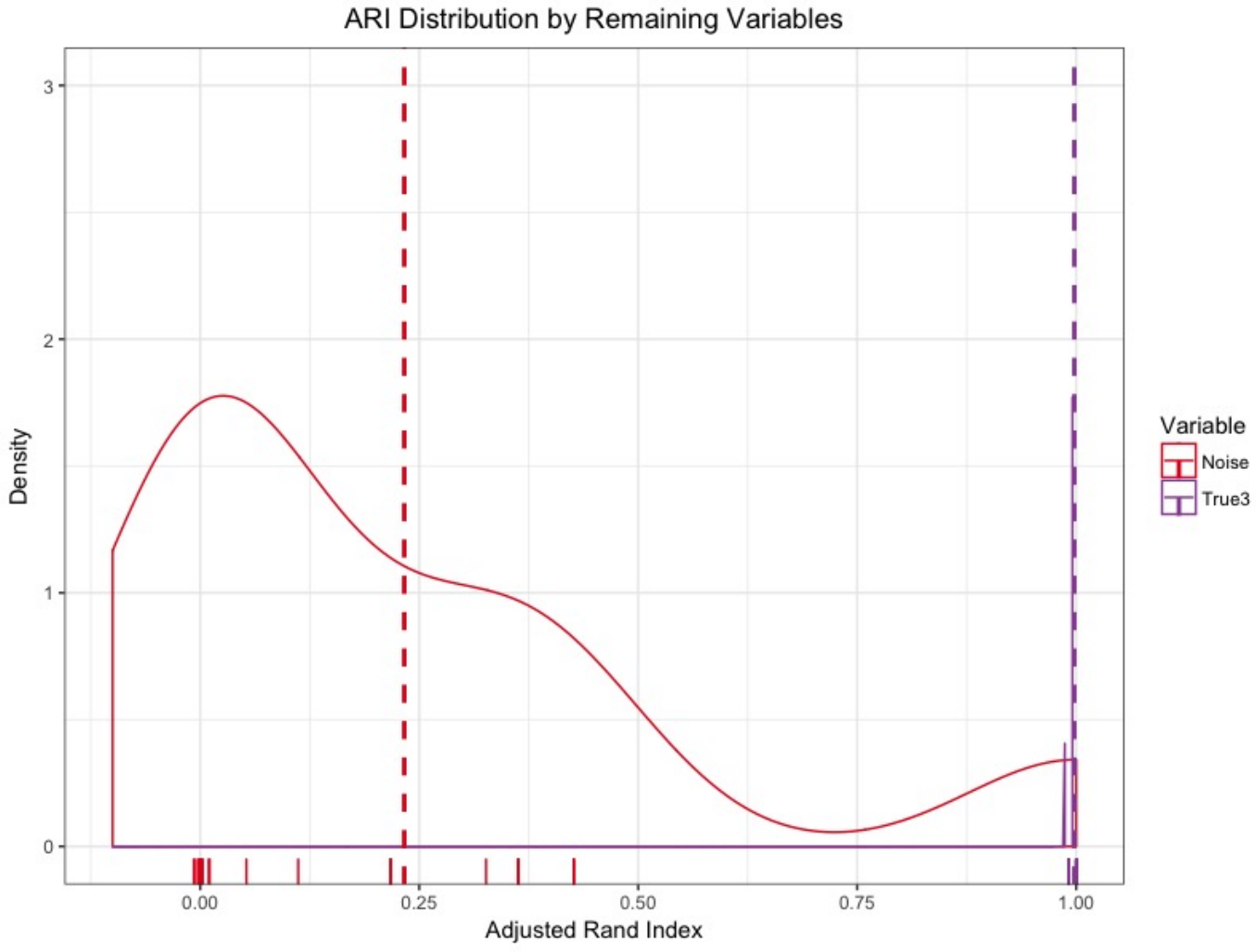




Identified 3 well separated clusters

Given {True4,True1,True2} Most Consistent: True3

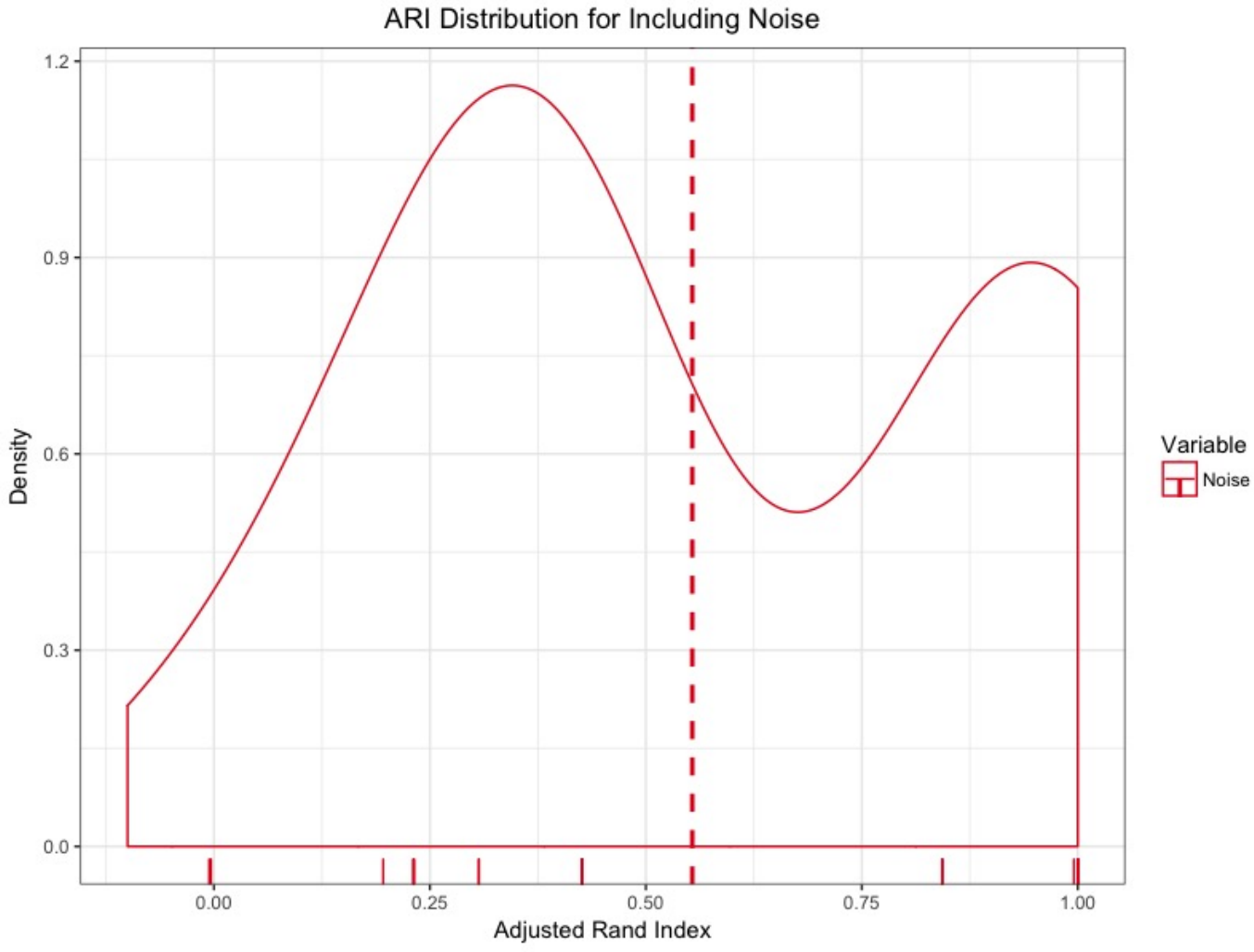
Still consistent results, $\overline{ARI}_{\{True4, True1, True2, True3\}} = 0.9979$



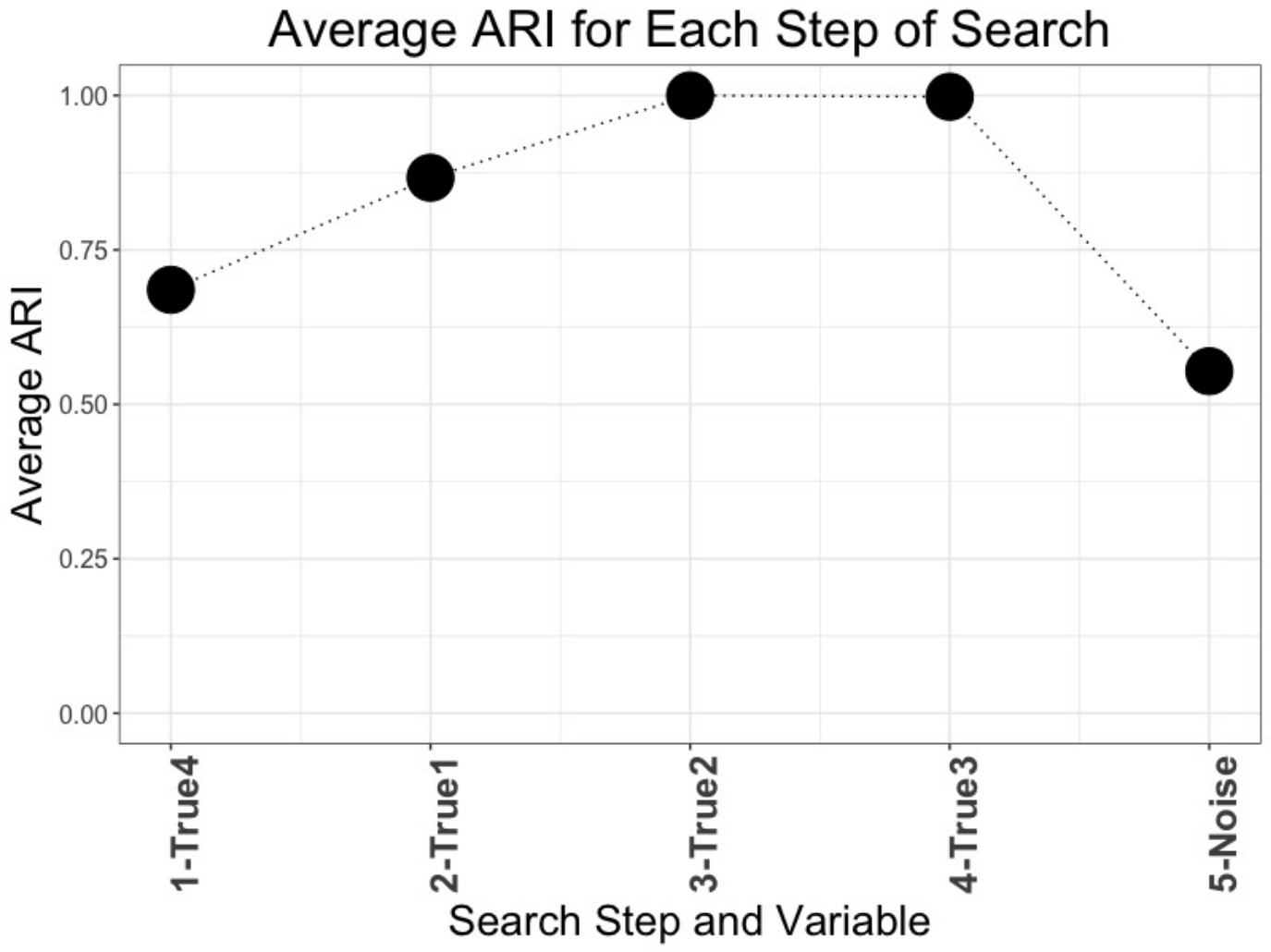
What About Including Noise?

Severe drop in consistency,

$\overline{ARI}_{\{True4, True1, True2, True3, Noise\}} = 0.5537$



Summary of Demo Search



Clearly see the drop from including Noise

What About The Number Of Clusters?

Only considered the true number of clusters, but...

What About The Number Of Clusters?

Only considered the true number of clusters, but...

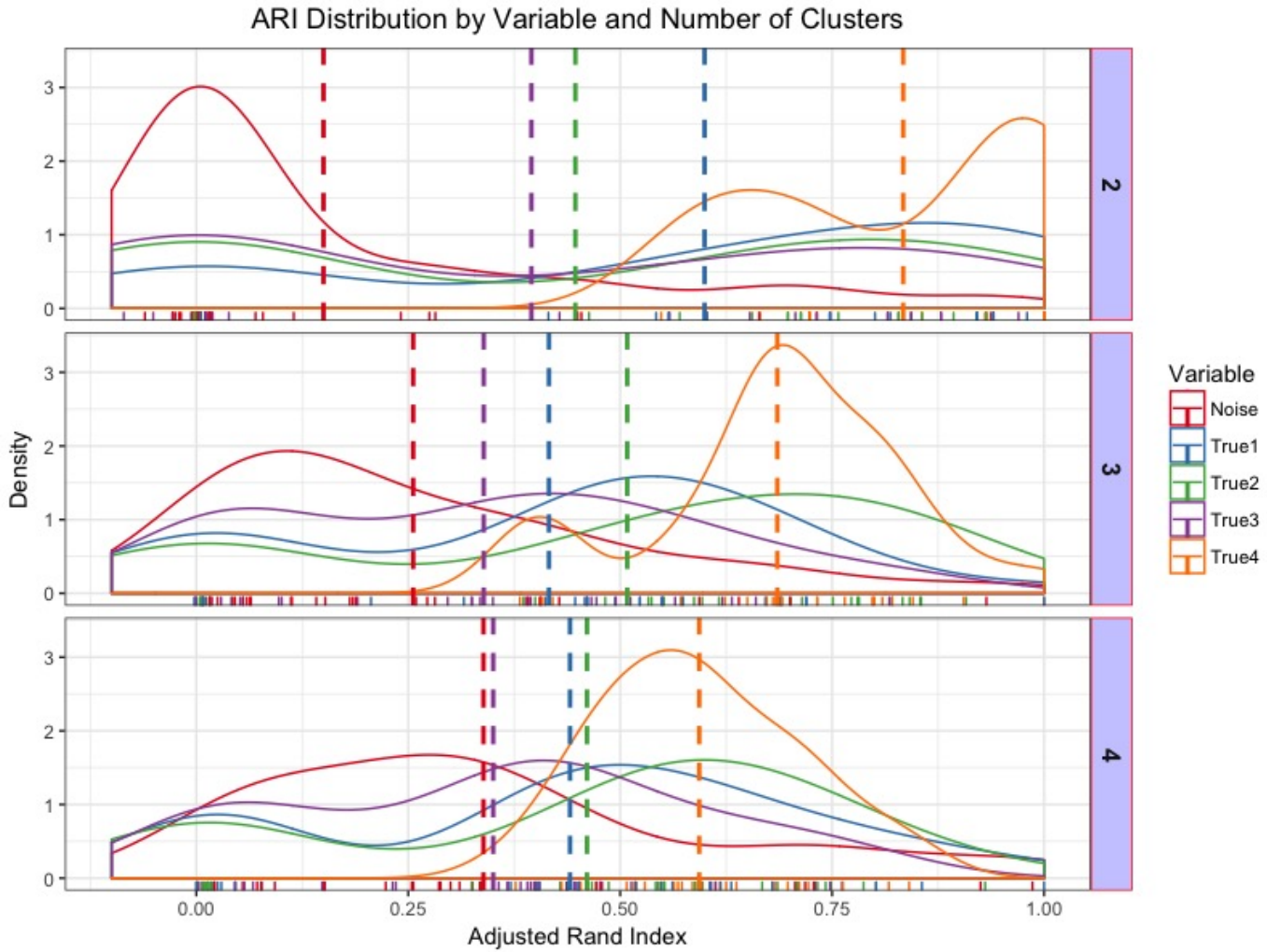
- $k = 1, 2, \ldots, K$ number of clusters
- $p_{m,k,S}$ defines a partition with method m and k clusters, given set of variables S
- $\overline{ARI}_{k,S} = \frac{1}{\binom{|M|}{2}} \sum_{i \neq j} ARI(p_{m_i,k,S},\ p_{m_j,k,S})$ measures the consistency for set of variables S with k clusters

Search over number of clusters and variables to identify the combination consistently agreeing on cluster structure

Most Consistent Combo: k=2, True4

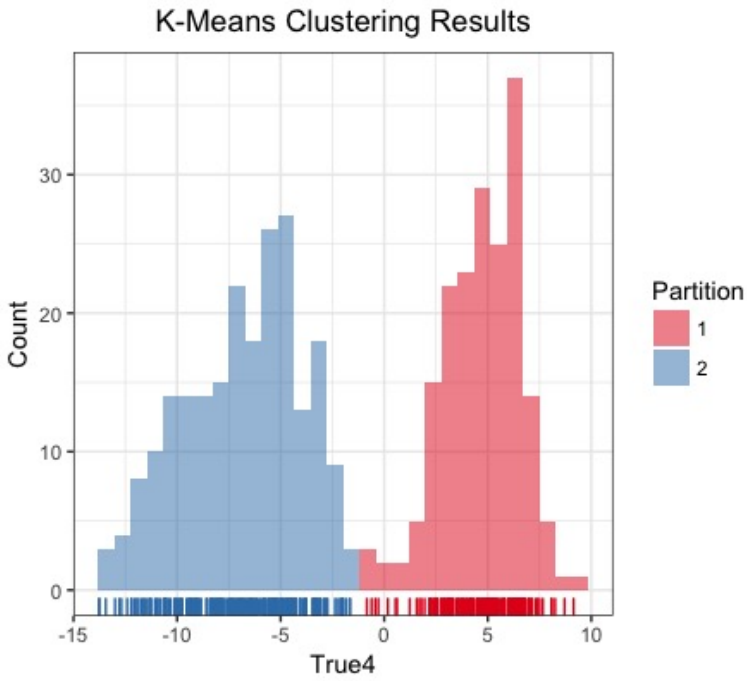
Still True4, but more consistent with 2 clusters

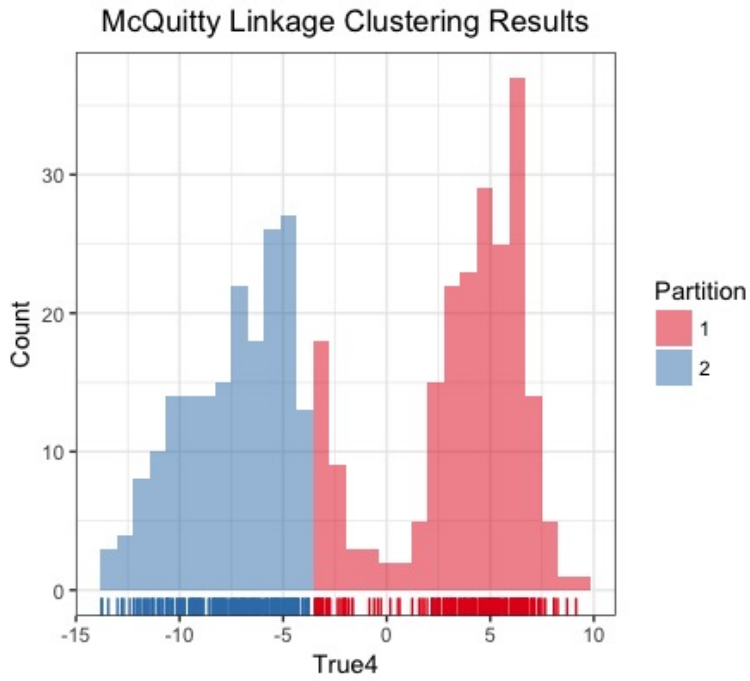
$\overline{ARI}_{2,\{True4\}} = 0.8338$



Consistent Clustering With 2 Clusters

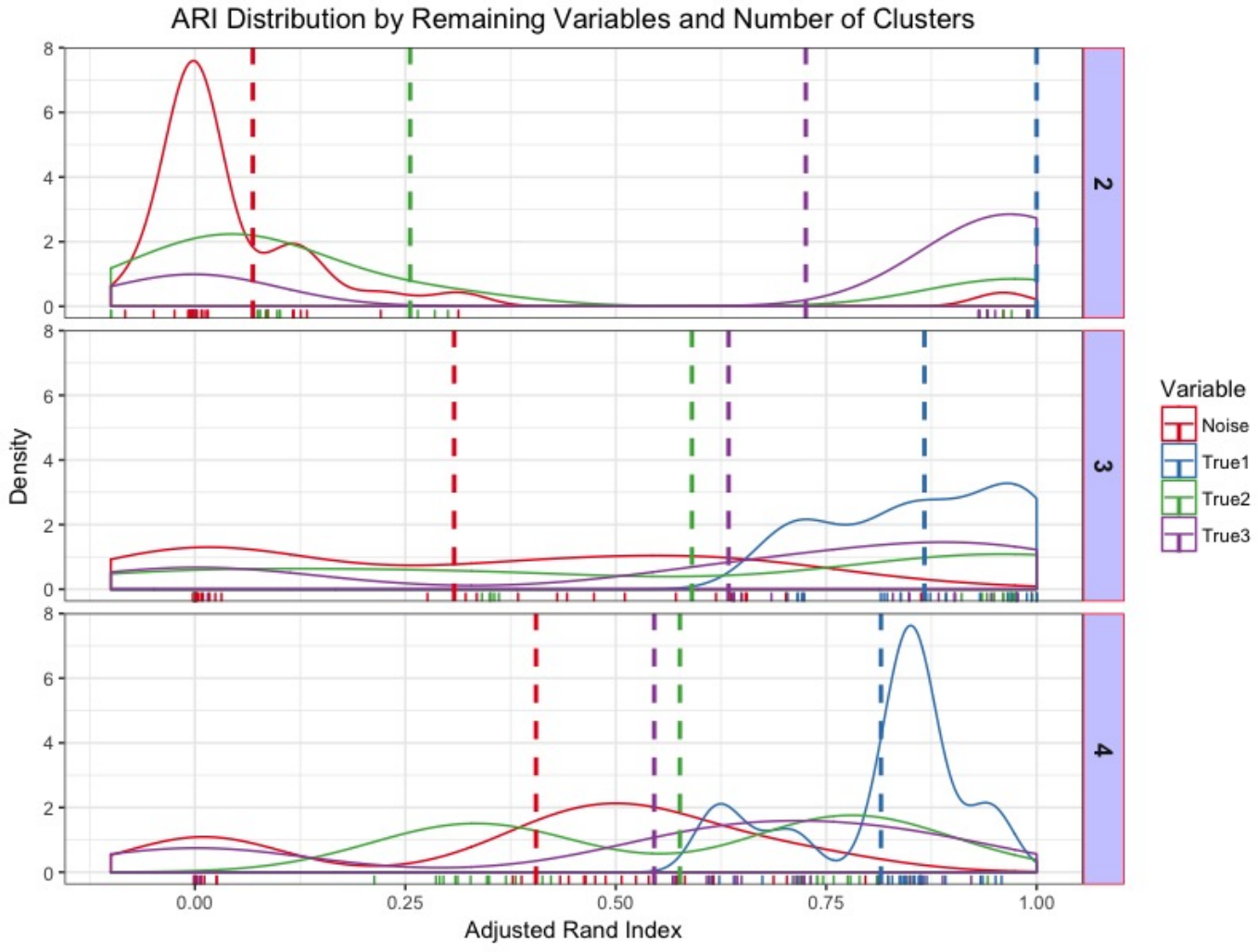
More consistent due to bimodal nature of True4





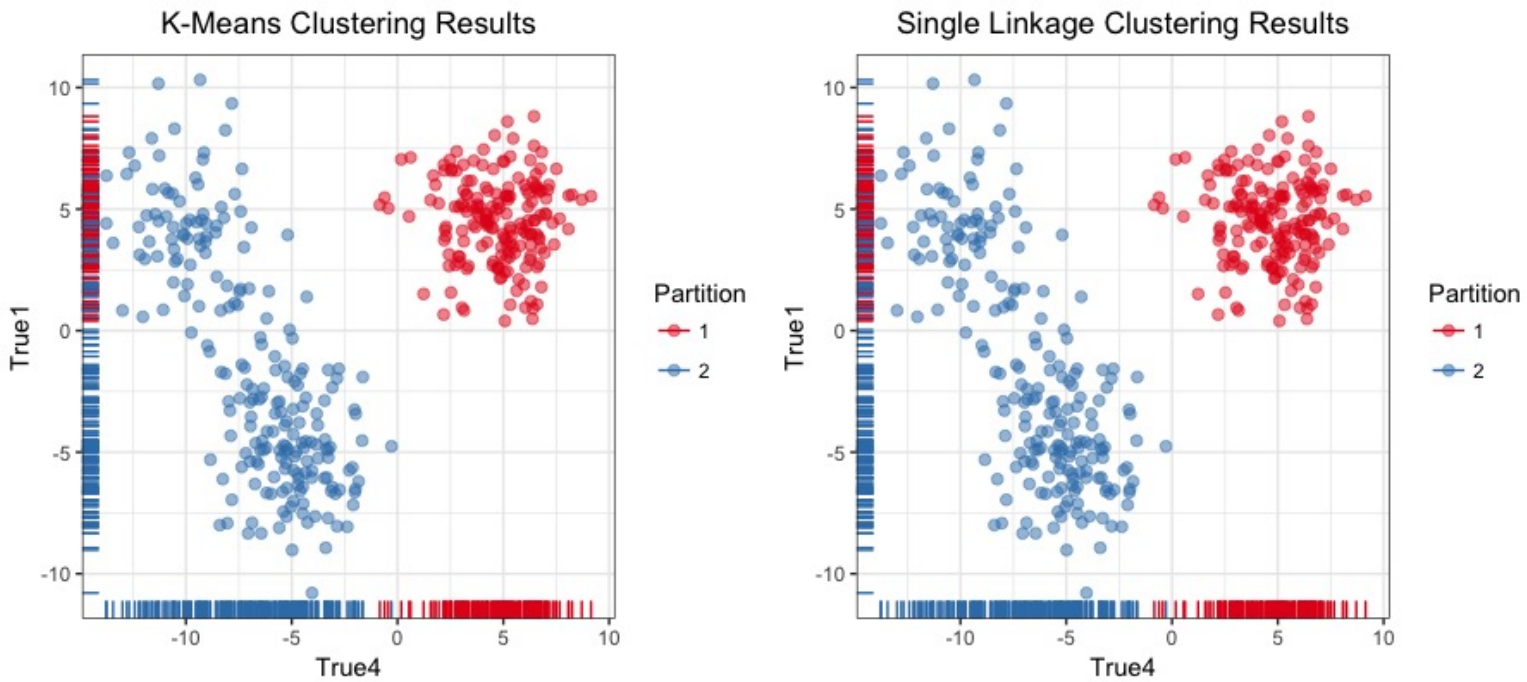
Given {True4}: k=2, True1

Perfect agreement, $\overline{ARI}_{2,\{True4,True1\}} = 1$



Consistent Clustering With 2 Clusters

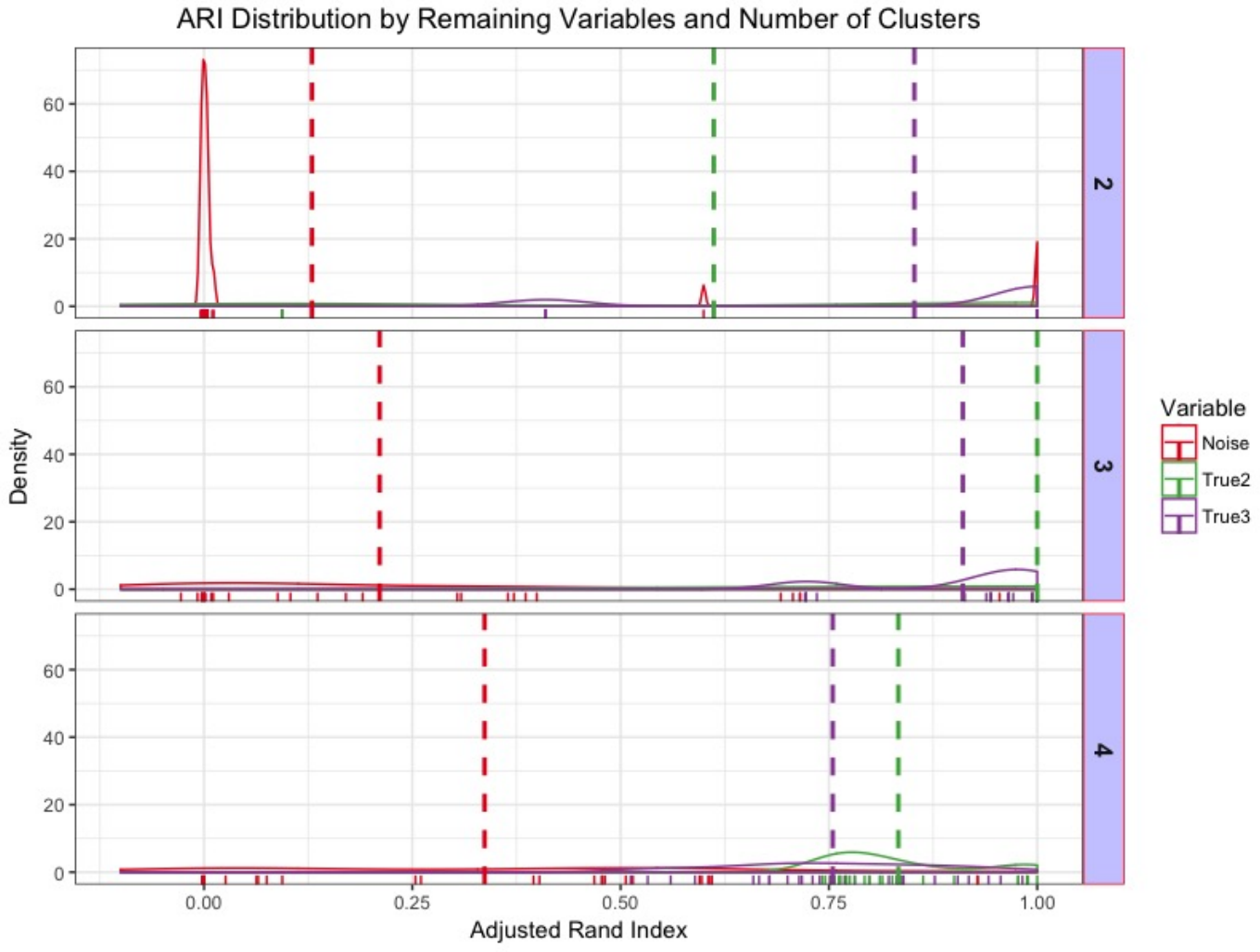
$ARI(p_{m_1,2,\{True4,True1\}},\ p_{m_2,2,\{True4,True1\}}) = 1,$
across all 28 pairs of methods



Identified 2 well separated clusters

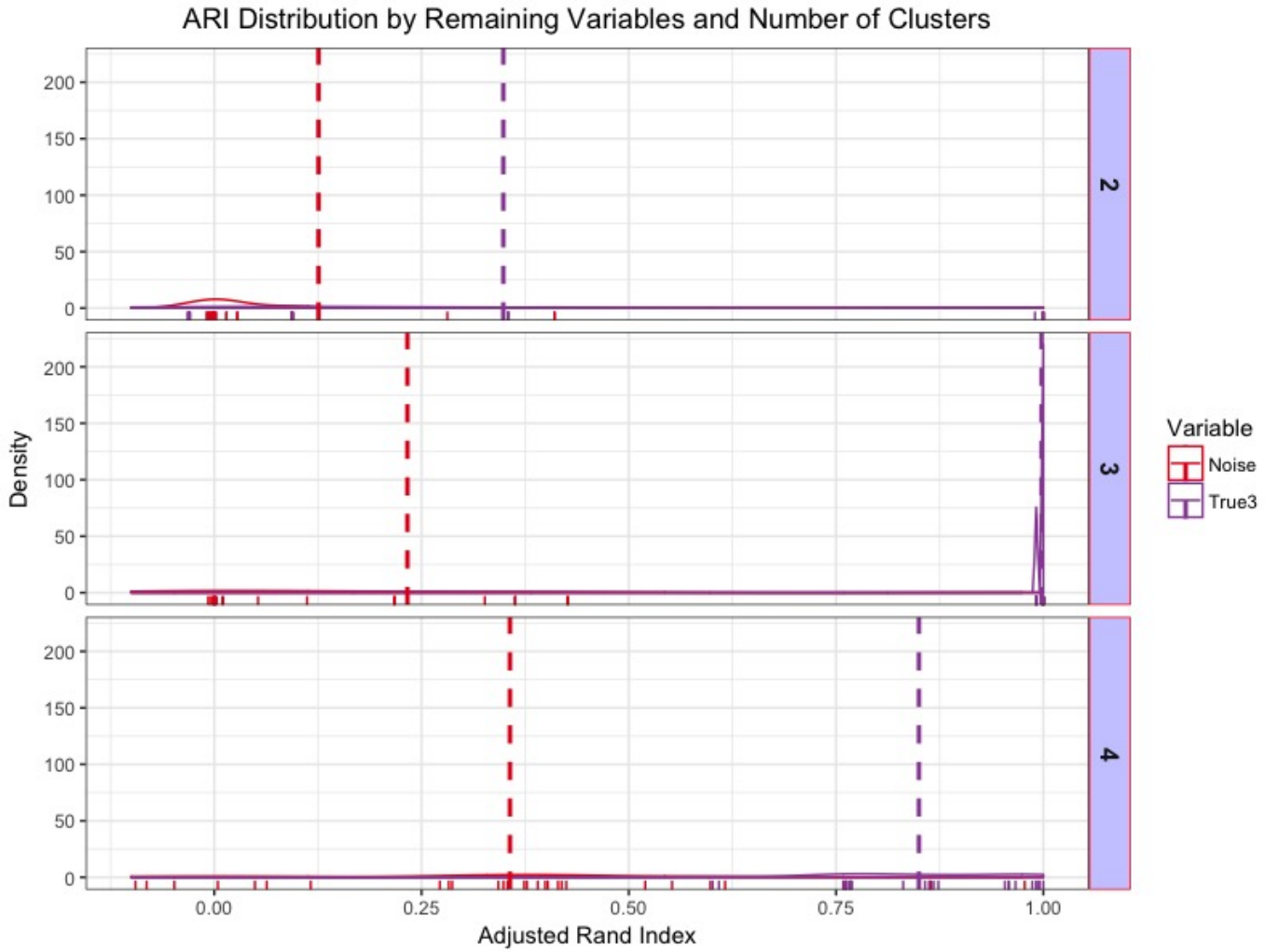
Given {True4, True1}: k=3, True2

Returns to previous result, $\overline{ARI}_{3,\{True4, True1, True2\}} = 1$



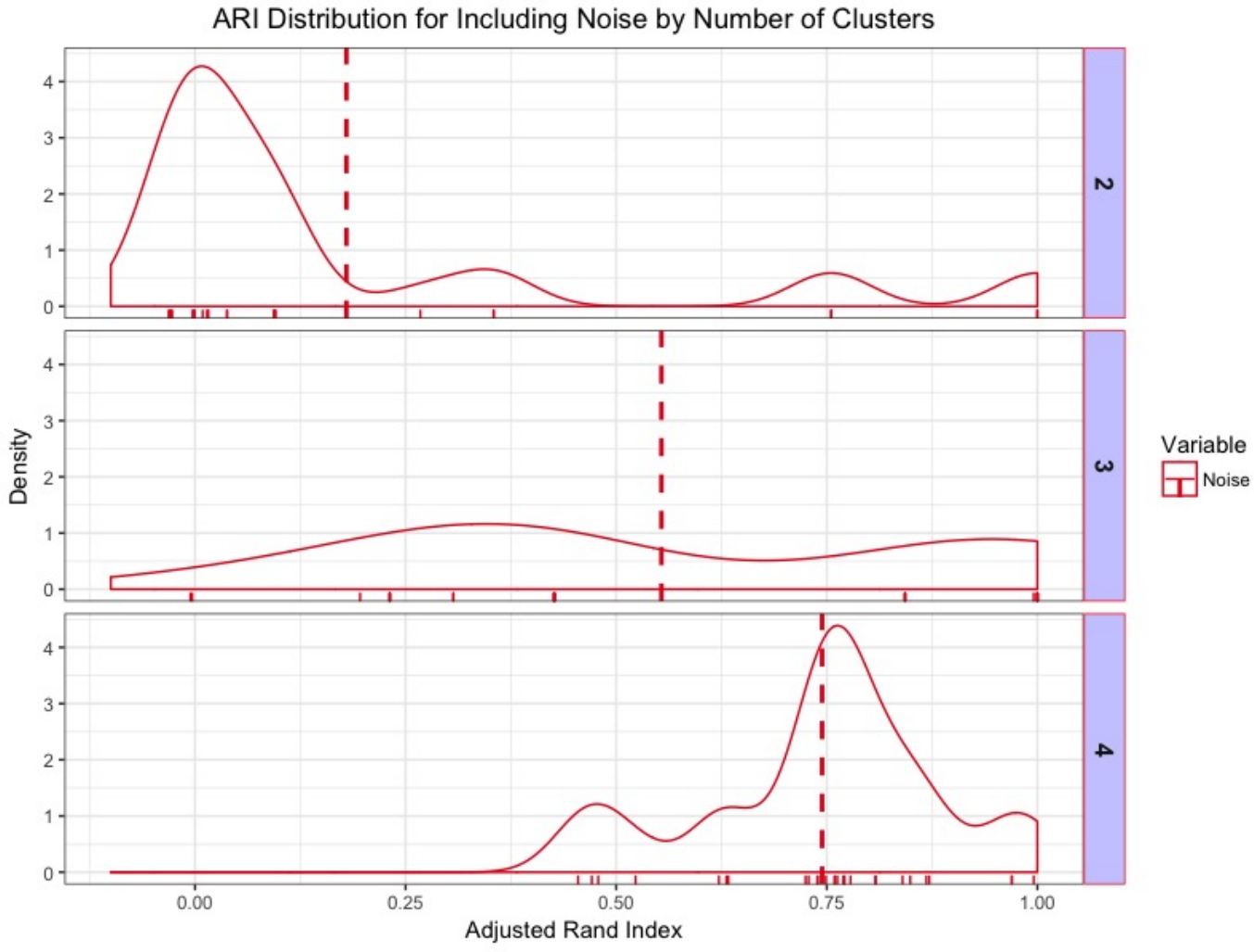
Given {True4, True1, True2}: k=3, True3

Again the same result, $\overline{ARI}_{3,\{True4, True1, True2, True3\}} = 0.9979$

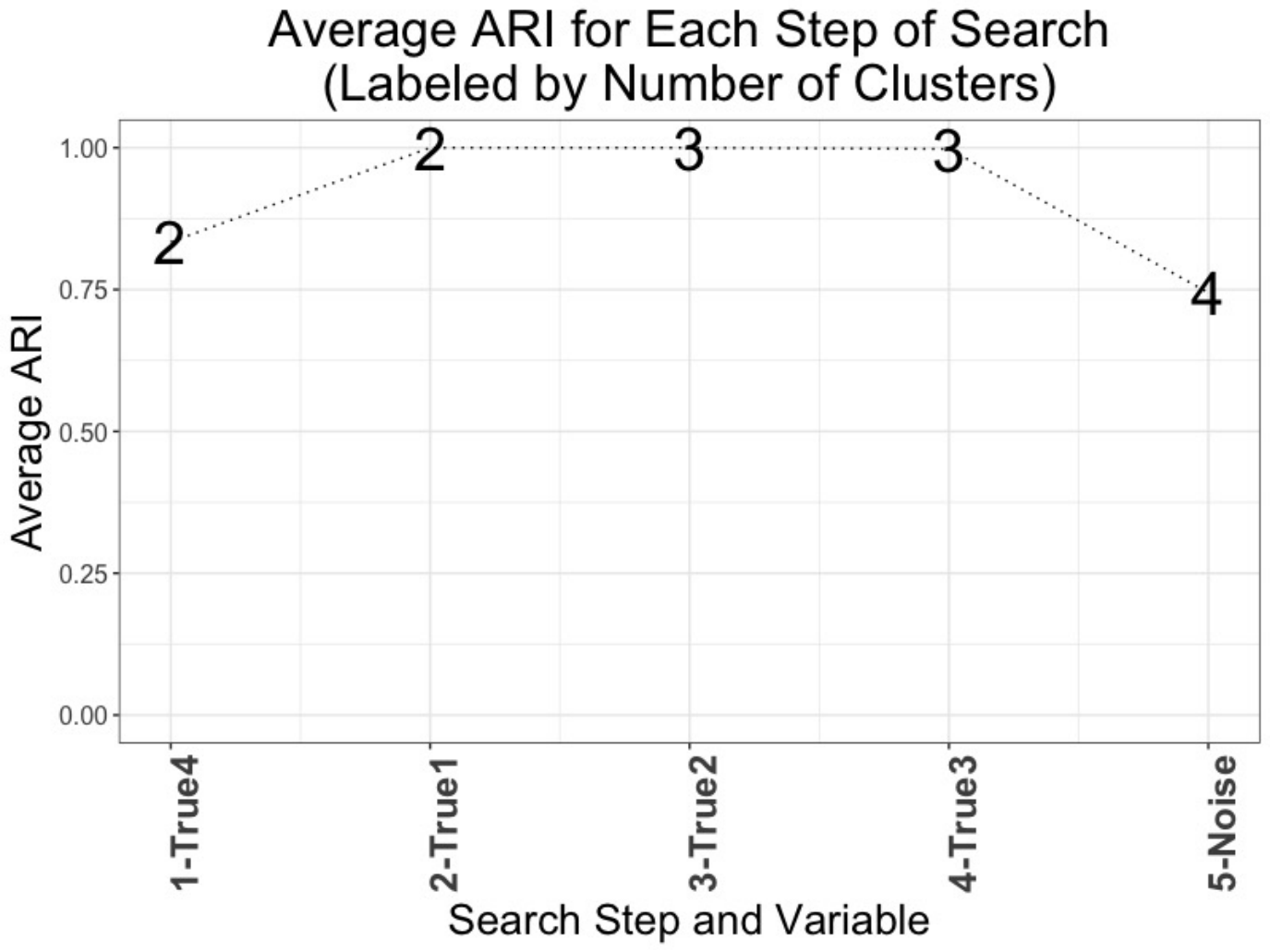


What About Including Noise?

Drop in consistency, $\overline{ARI}_{4,\{True4, True1, True2, True3, Noise\}} = 0.7444$

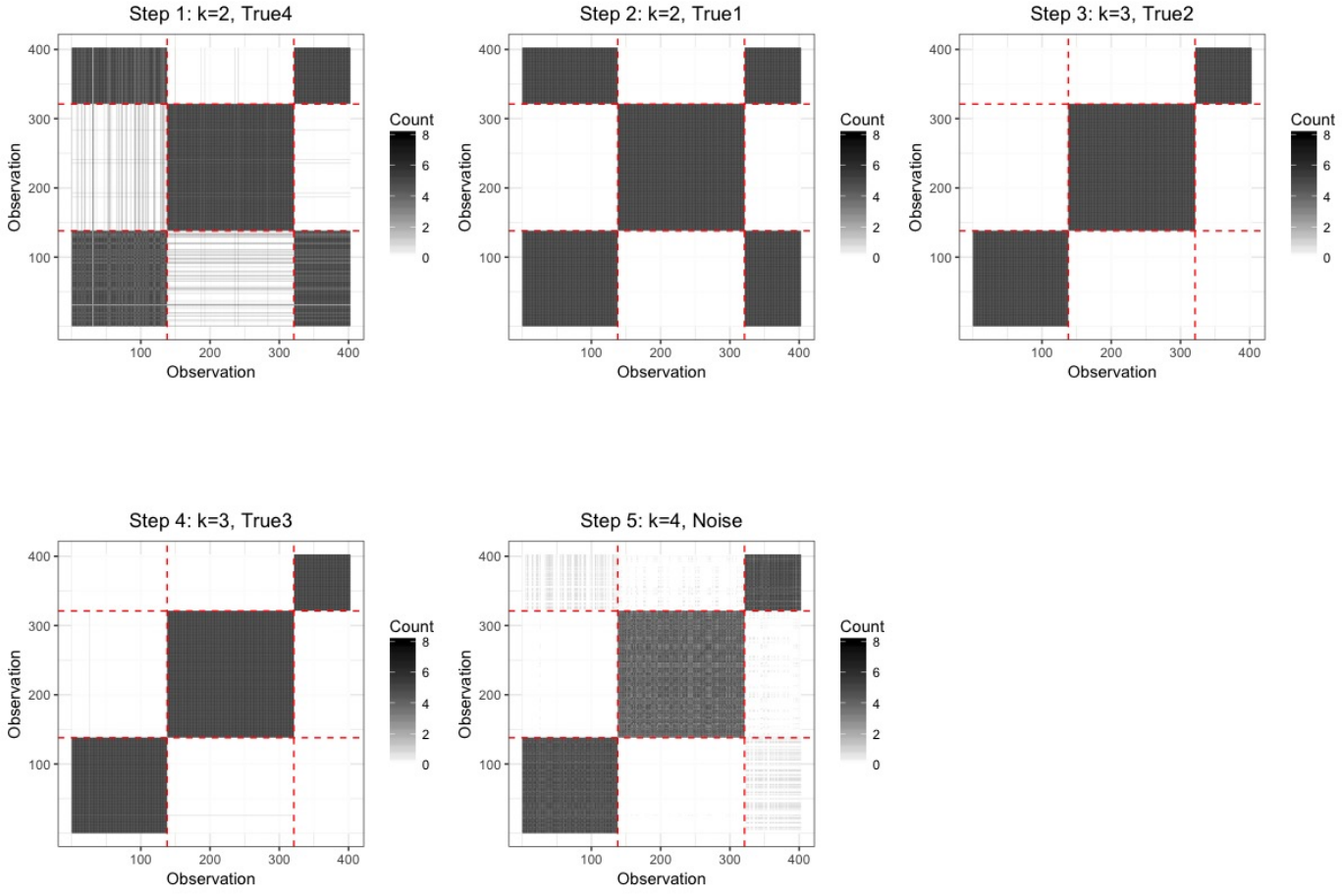


Summary of Demo Search



Clearly see the drop from including Noise

Summary of Demo Search



Stopping Criteria?

Perfect Agreement:

- $\overline{ARI}_{k,S\cup\{d^*\}} = 1$

Minimum Threshold:

- Only include d^* if $\overline{ARI}_{k,S\cup\{d^*\}} > ARI_{min}$

Percentage Drop:

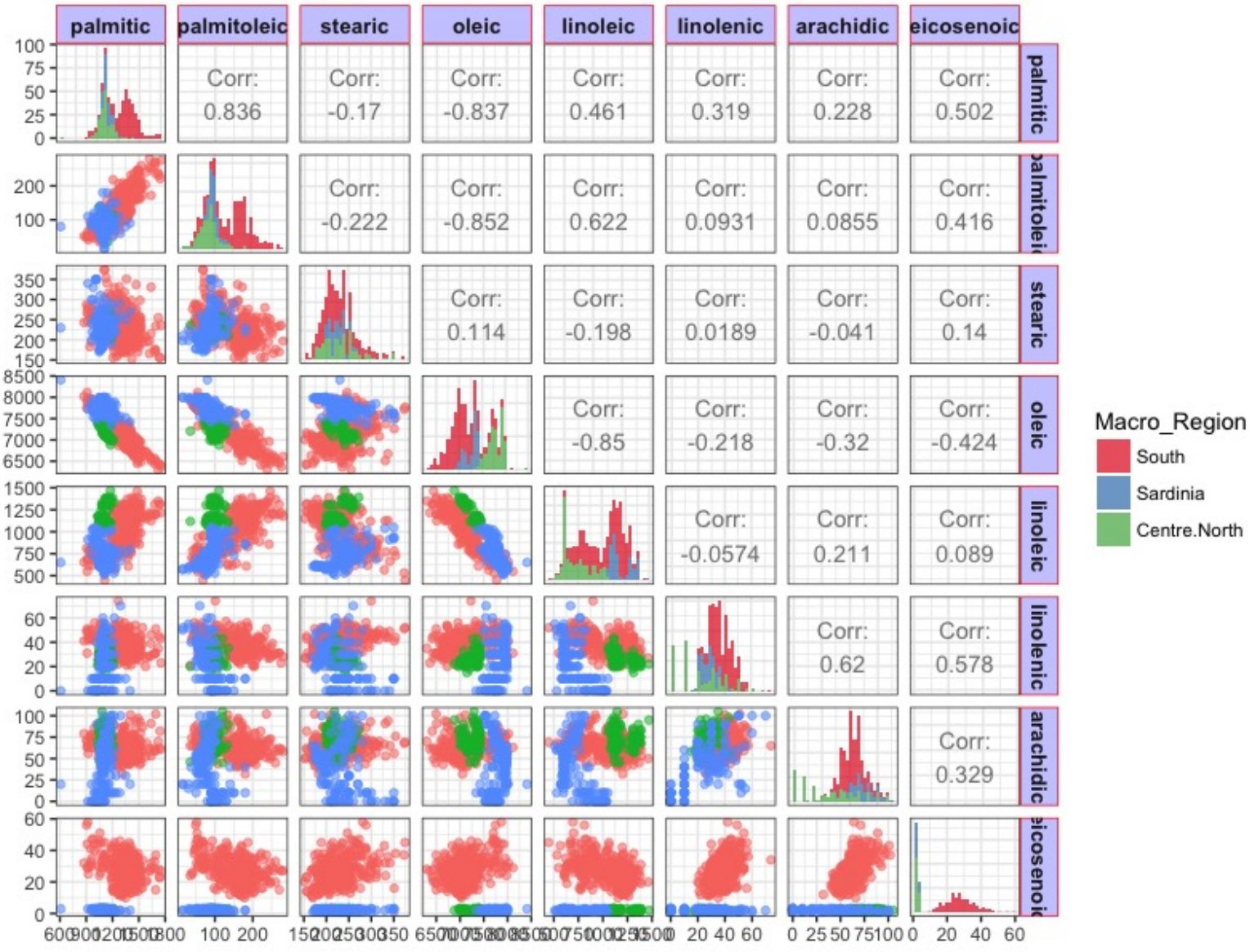
- Only include d^* if $\overline{ARI}_{k,S\cup\{d^*\}} > C * \overline{ARI}_{k,S}$, where C is a percentage

User can input desired level of consistency

Olive Oil Data

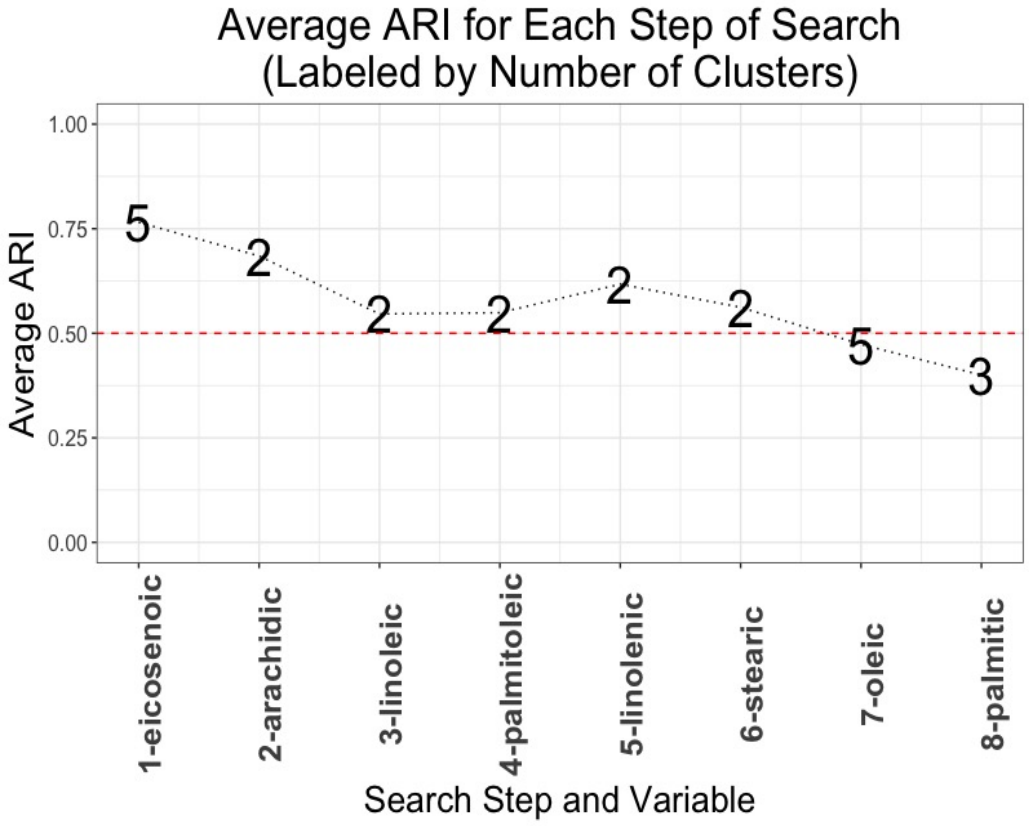
- 572 observations, each a different type of olive oil
- 8 variables for chemical measurements of acids
- 2 known labels:
  - Macro-Regions: 3 groups
  - Sub-Regions: 9 groups

Oils by Macro-Region



Macro-Region Consistency Search

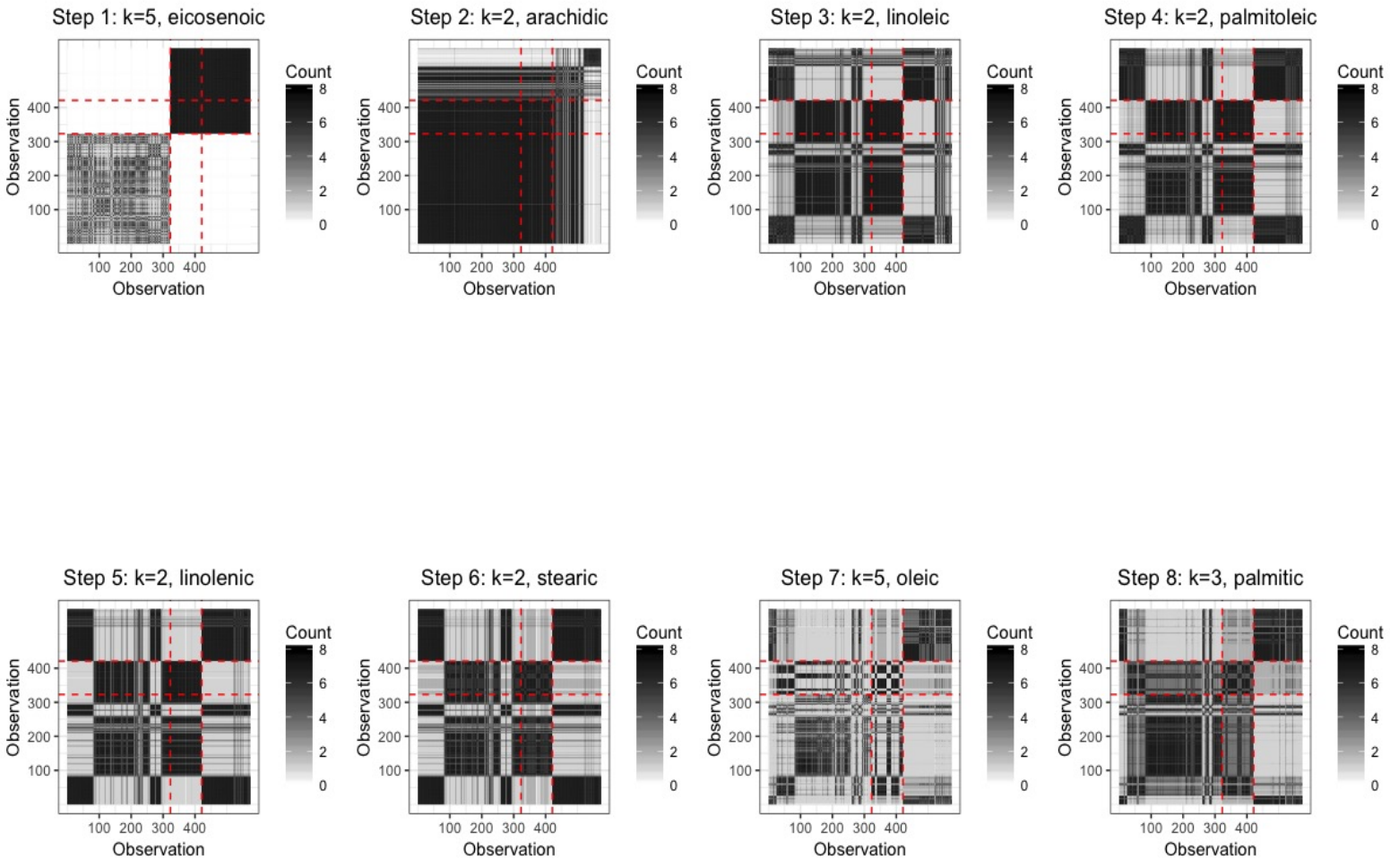
Search over $k = 2, \ldots, 5$: stop after Step 5, $\overline{ARI}_{2,S} = 0.6174$



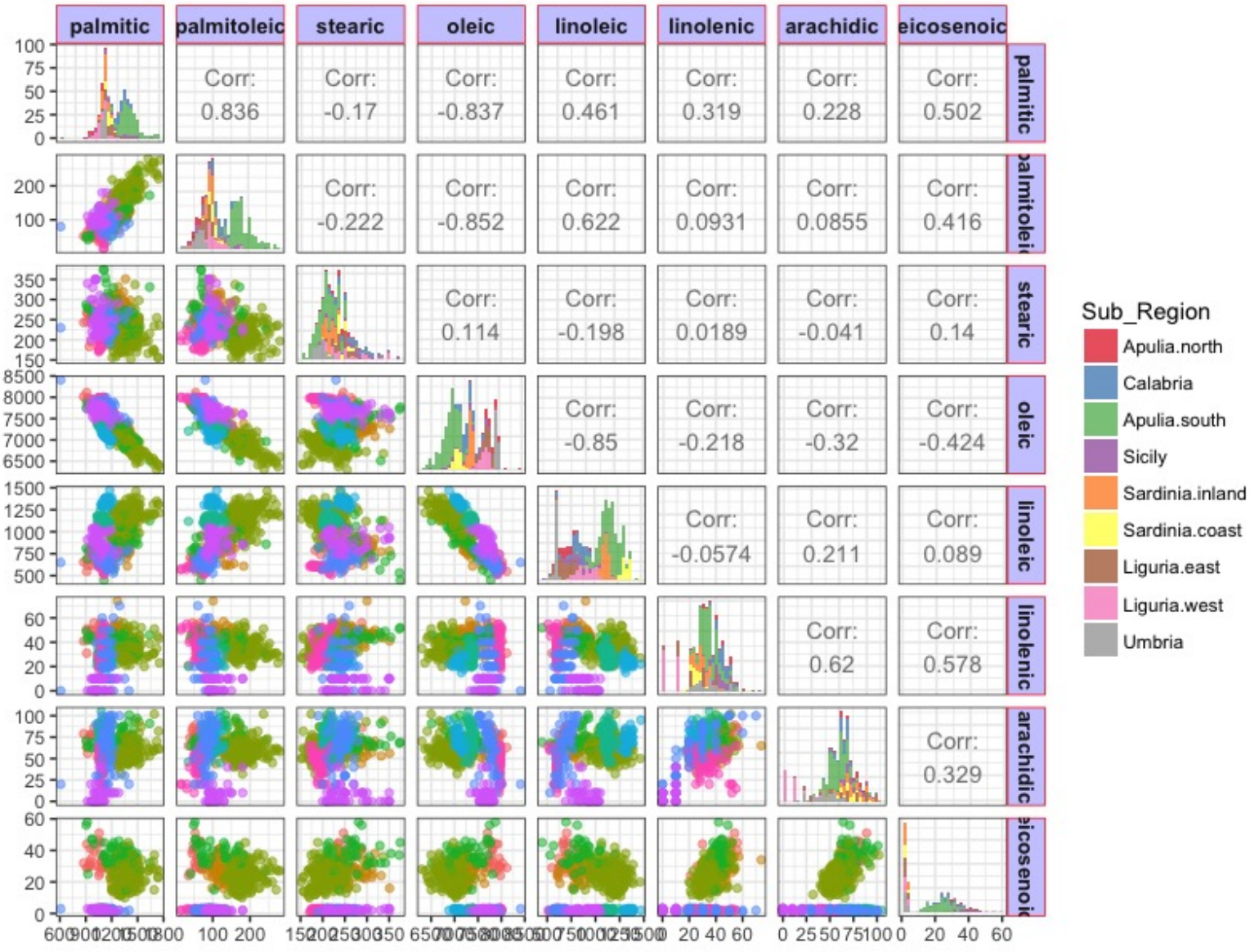
Model-based clustering, with `clustvarsel` *(Raftery & Dean, 2006)*:

Chose all 8 variables, with $k = 5$ (while $\overline{ARI}_{5,S \cup U} = 0.3955$)

Macro-Region Consistency Search

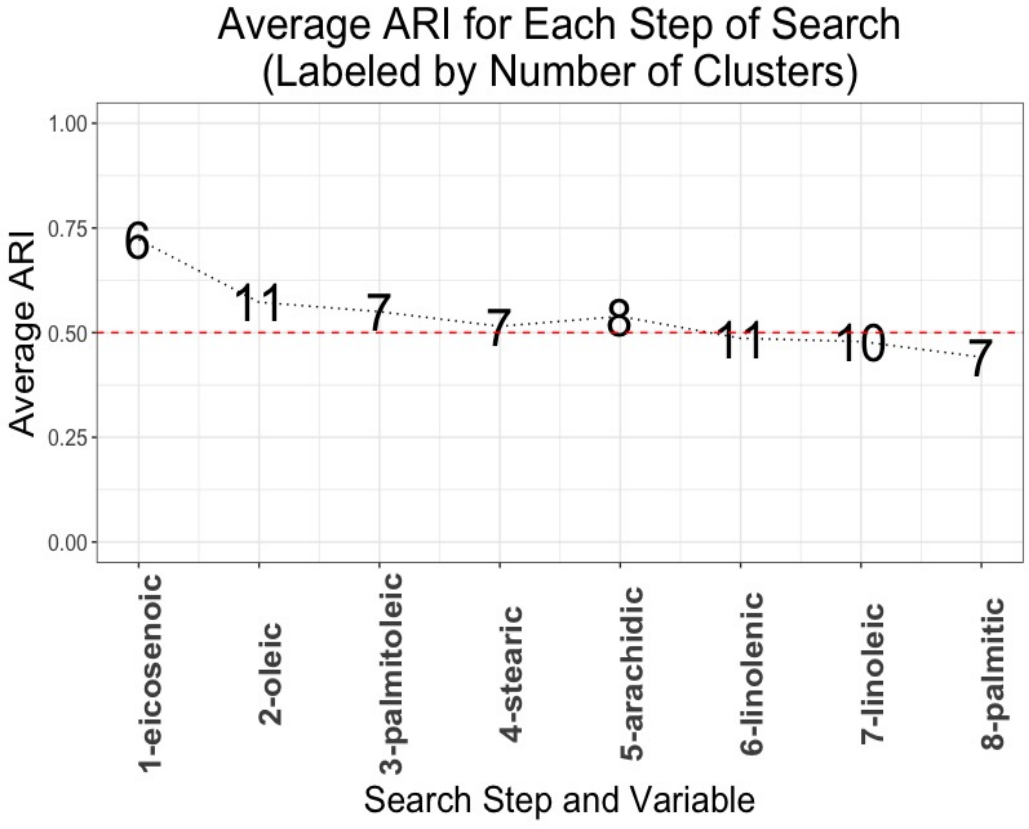


Oils by Sub-Region



Sub-Region Consistency Search

Search over $k = 6, \ldots, 12$: stop after Step 5, $\overline{ARI}_{8,S} = 0.5457$



Model-based clustering:

Chose 7 variables with $k = 9$, didn't select oleic

Future Work

- Include a removal step in the search
- Only considered average ARI, but distributions are multimodal and assymetrical
- Weighting of comparisons,
  e.g. less weight on single vs complete partitions
- Sensitive to methods considered, expand with concept

- Interpretation of changing clusters with different methods
- Mode hunting for specific clustering methods
- Rather than stopping search, consider full sequence path

References I

[Hastie et al., 2009]. [Brusco and Cradit, 2001].
[Carmone et al., 1999].
[Albatineh and Niewiadomska-Bugaj, 2011].
[Fraley and Raftery, 1998]. [Raftery and Dean, 2006].
[Hubert and Arabie, 1985]. [Milligan, 1996].
[Brusco and Steinley, 2007].


Albatineh, A. N. and Niewiadomska-Bugaj, M. (2011).
Mcs: A method for finding the number of clusters.
Journal of Classification, 28:184–209.


Brusco, M. and Cradit, J. D. (2001).
A variable-selection heuristic for k-means clustering.
Pyschometrika, 66(2):249–270.

References II


Brusco, M. and Steinley, D. (2007).
A comparison of heuristic procedures for mininum within-cluster sums of squares partitioning.
Pyschometrika, 72:583–600.


Carmone, F. J., Kara, A., and Maxwell, S. (1999).
Hinov: A new model to improve market segmentation by identifying noisy variables.
Journal of Marketing Research, 36:501–509.


Fraley, C. and Raftery, A. E. (1998).
How many clusters? which clustering method? answers via model-based cluster analysis.
The Computer Journal, 41(8):578–588.

References III


Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer, New York, New York.


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


Milligan, G. W. (1996).
Clustering validation: Results and implications for applied analyses.
In Arabie, P., Hubert, L. J., and Soete, G. D., editors, *Clustering and Classification*, pages 341–375. World Scientific, Singapore.

References IV


Raftery, A. E. and Dean, N. (2006).
Variable selection for model-based clustering.
Journal of the American Statistical Association, 101(473):168–178.

Maximum clustering similarity

MCS: A Method for Finding the Number of Clusters
(Albatineh & Niewiadomska-Bugaj, 2011)
Use corrected similarity indices to compare clustering methods

- Based on the number of pairs of data points that are (not) placed into the same cluster
- R (Rand, 1971), FM (Fowlkes and Mallows, 1983), K (Kulczynski, 1927)
- Adjusted Rand index (ARI) (Hubert and Arabie, 1985):

$$ARI = \frac{R - E(R)}{1 - E(R)}$$

- ARI close to 0 due to chance, max value of 1

Choose the number of clusters that is most frequently the maximum similarity value across all method comparisons

Variable selection with ARI

HINoV method (Carmone, Kara, & Maxwell, 1999)

- Begin with K-means partition, p_j, using only variable j for each of the $j \in \{1, \ldots, D\}$ variables
- Compute ARI_{jk} between each of the $\binom{D}{2}$ pairs of partitions, $k \in \{1, \ldots, D\}$ and $k \neq j$
- Rank variables by $TOPRI_j = \sum_{k=1}^{D} ARI_{jk}$

Variable selection with ARI

A Variable-Selection Heuristic for K-means Clustering
(Brusco & Cradit, 2001)

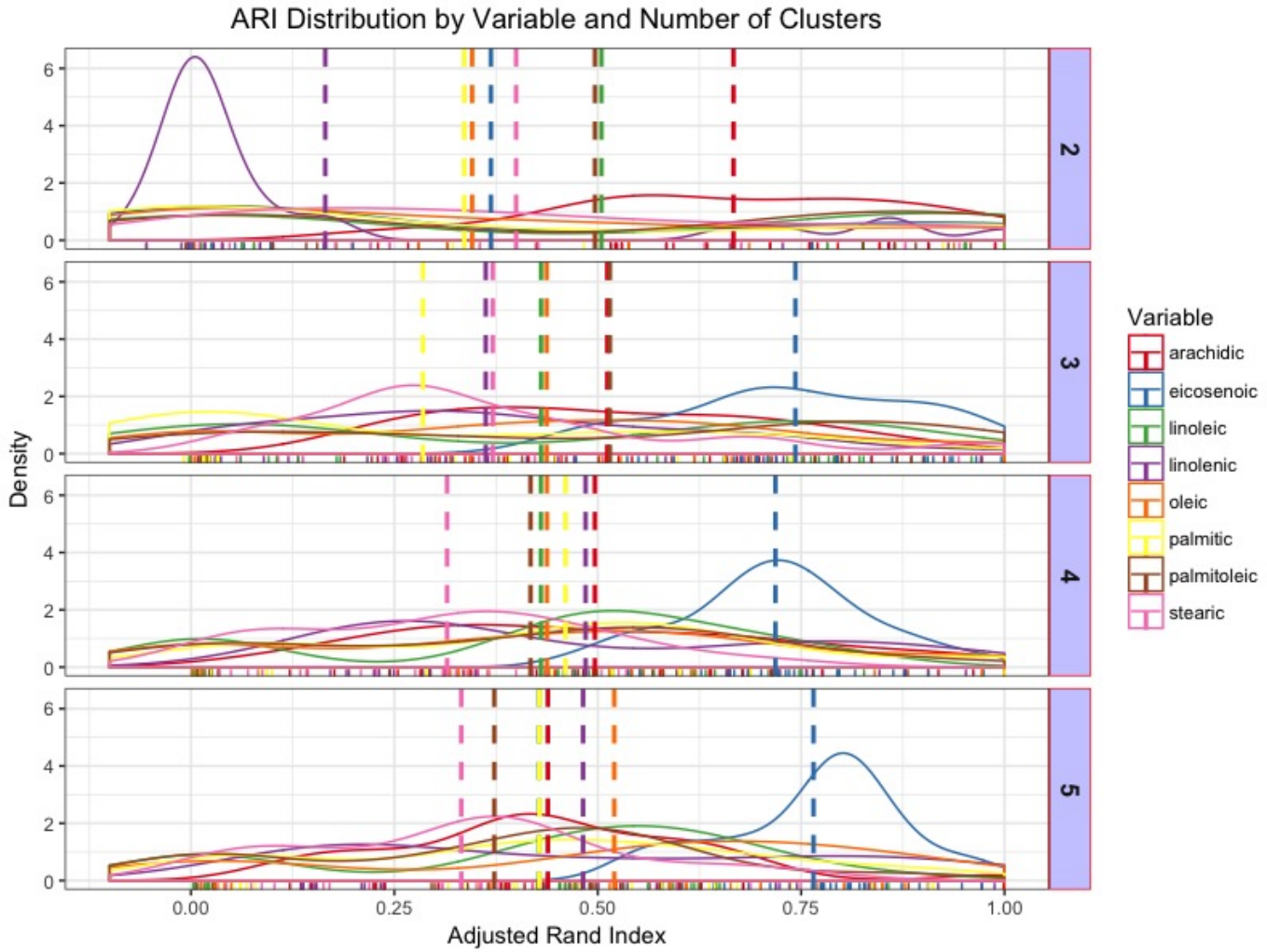
- Begin with K-means partition, p_j, using only variable j for each of the $j \in \{1, \ldots, D\}$ variables
- Compute ARI_{jk} between each of the $\binom{D}{2}$ pairs of partitions, $k \in \{1, \ldots, D\}$ and $k \neq j$
- Create partition w_{jk} with variables j and k, compute ratio of the between cluster sum-of-squares to the total sum-of-squares for partition
- Select variables with highest ratio given ARI_{jk} is above a threshold
- Calculate ARI between w_{jk} and $p_{j'}$ for each of the remaining variables $j' \in \{1, \ldots, D\}$ and $j' \neq j, k$
- Select variable with highest ARI above threshold, repeat until ARI drops below this threshold

Olive Oil Demo

Search for smaller number of clusters similar to Macro

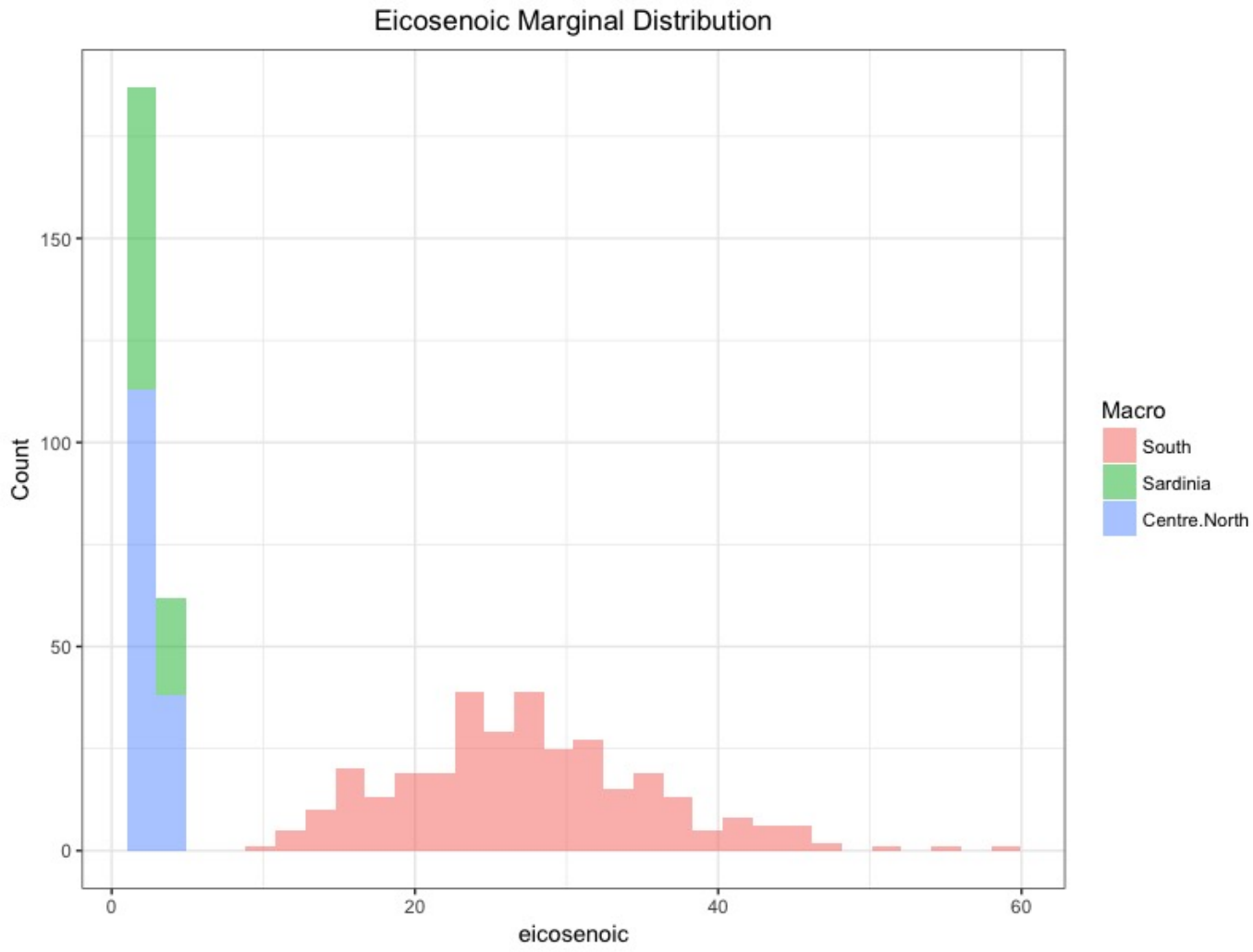
Most Consistent Combo: k=5, eicosenoic

Searching over 2-5 clusters, $\overline{ARI}_{\{eicosenoic\}} = 0.7652$



Why 5 Clusters?

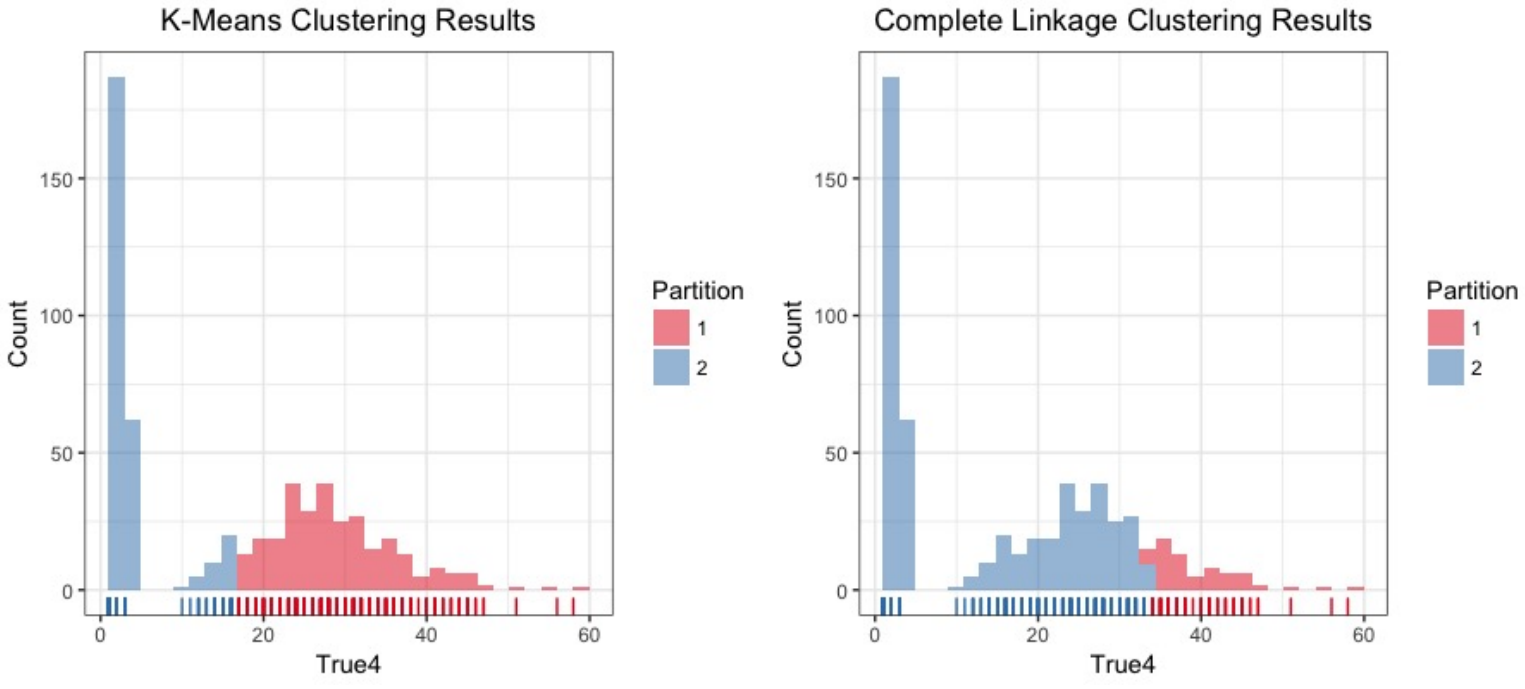
South appears to be separated from the other macro areas



Why 5 Clusters?

Not enough separation for 2 clusters,

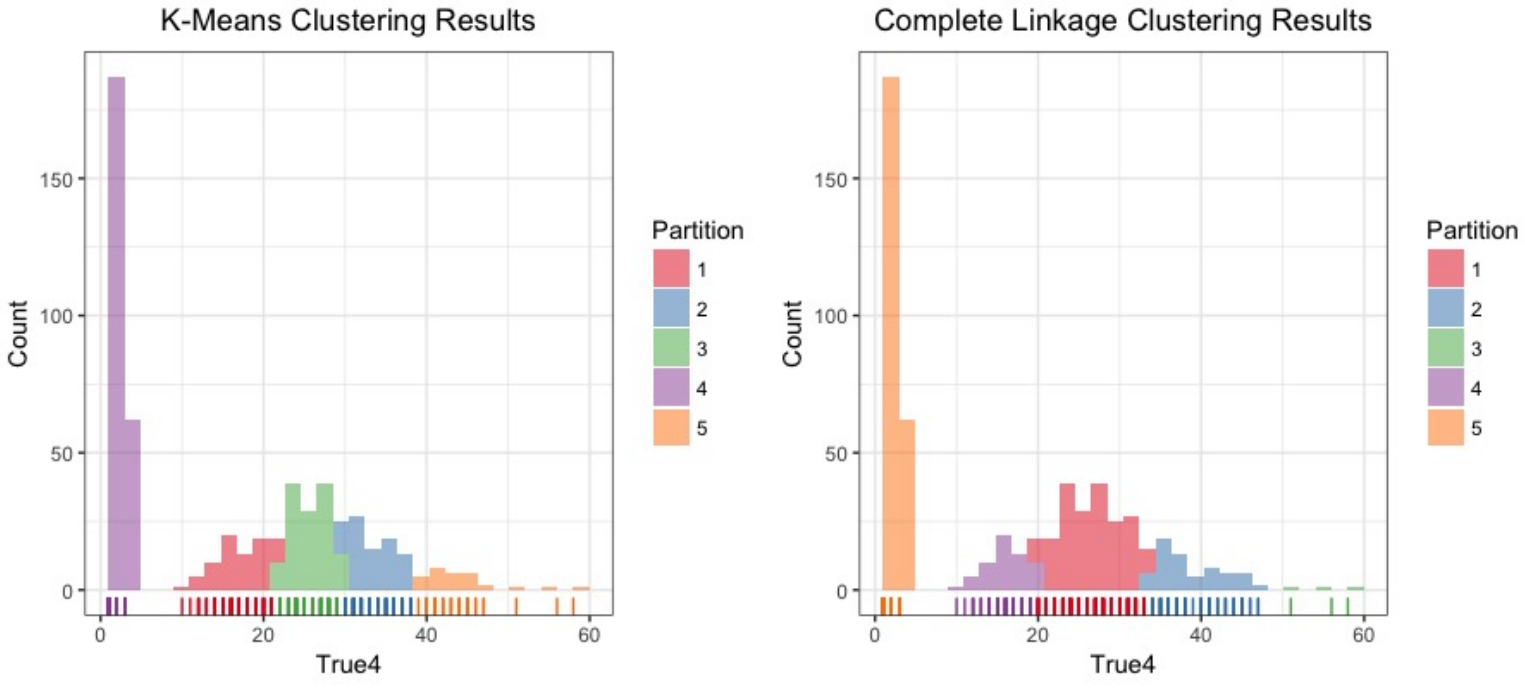
$ARI(p_{kmeans,2,\{eicosenoic\}}, p_{complete,2,\{eicosenoic\}}) = 0.0542$



Why 5 Clusters?

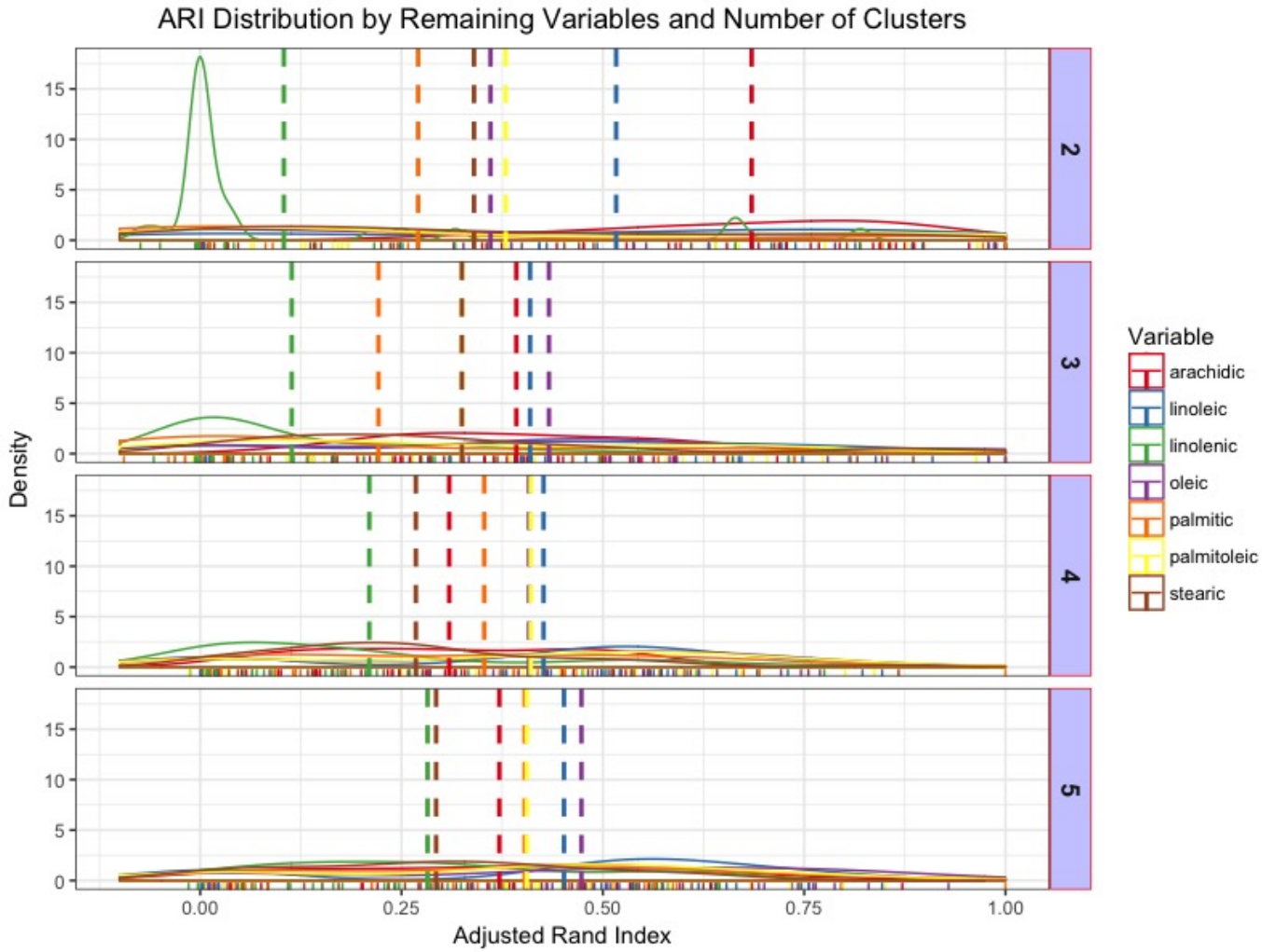
Consistent results with 5 clusters,

$ARI(p_{kmeans,5,\{eicosenoic\}}, p_{complete,5,\{eicosenoic\}}) = 0.8069$



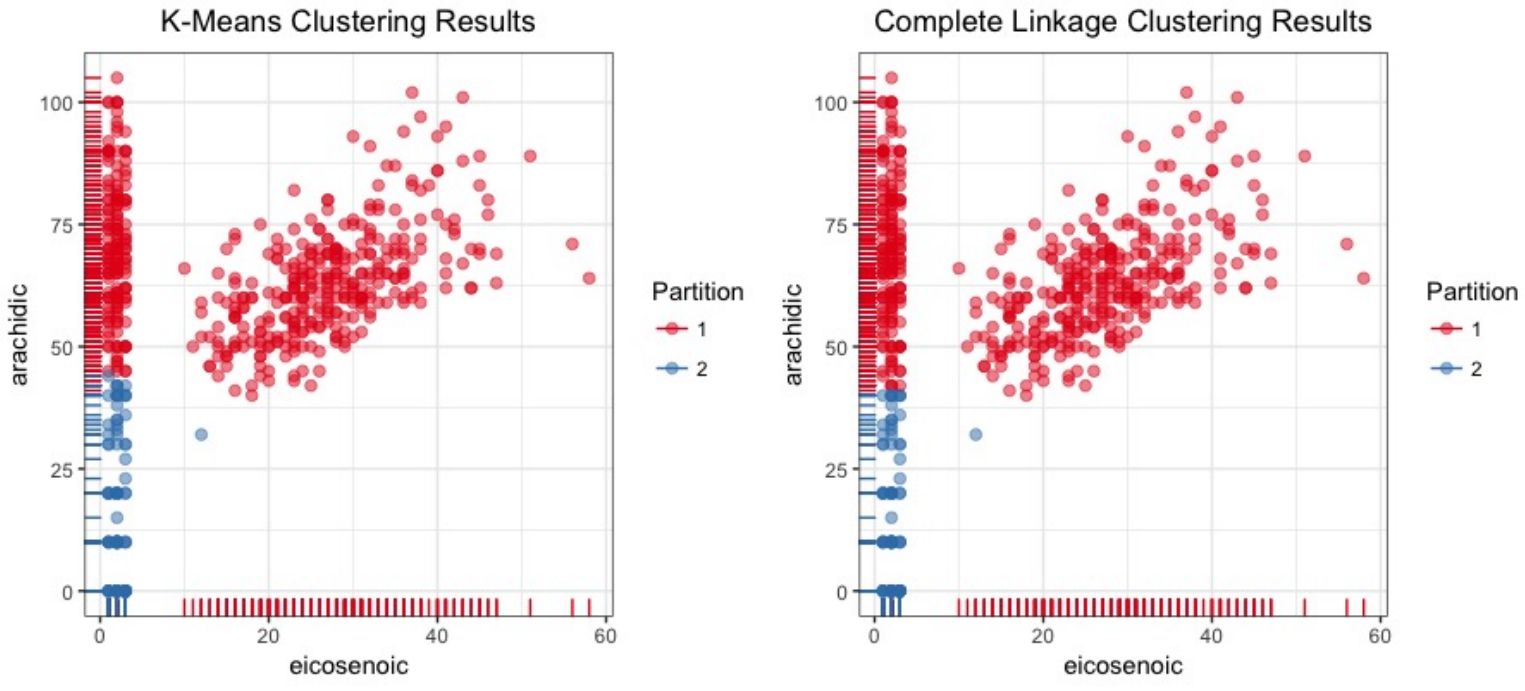
Given {eicosenoic}: k=2, arachidic

Switches to 2 clusters, $\overline{ARI}_{2,\{eicosenoic,arachidic\}} = 0.6848$



Why 2 Clusters?

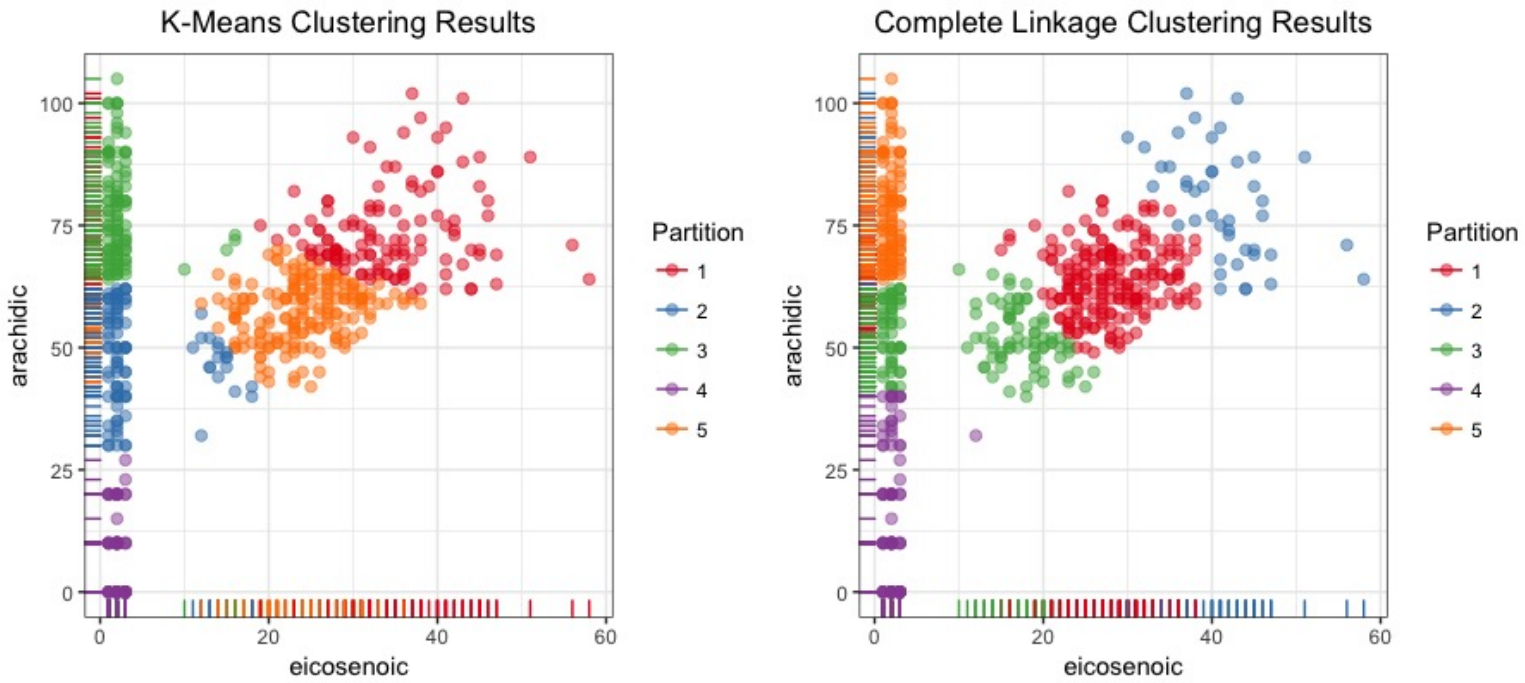
$ARI(p_{kmeans,2,\{eicosenoic,arachidic\}}, p_{complete,2,\{eicosenoic,arachidic\}}) = 0.9548$



Consistent with 2, but not well separated clusters

Why 2 Clusters?

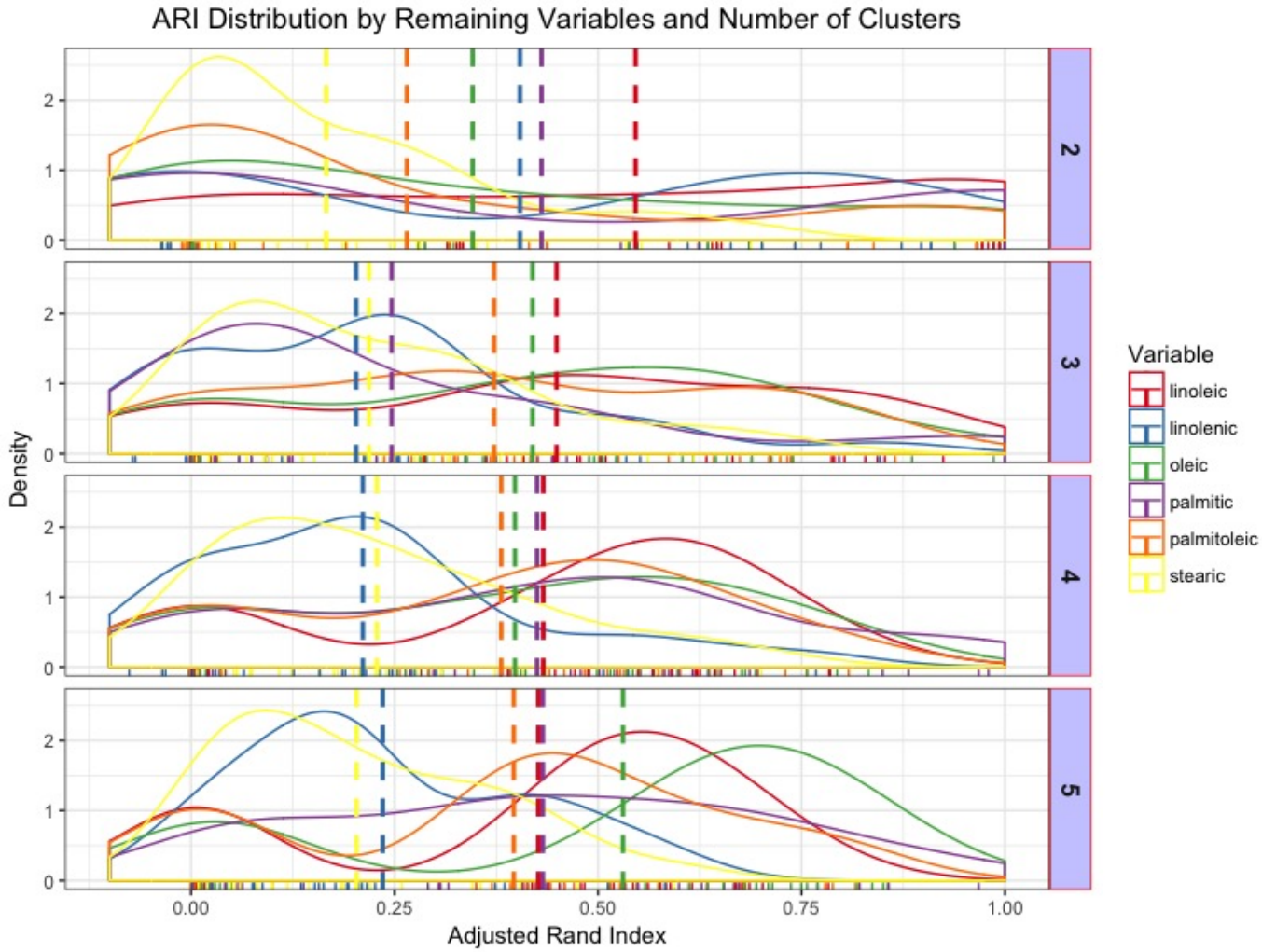
$ARI(p_{kmeans,5,\{eicosenoic,arachidic\}}, p_{complete,5,\{eicosenoic,arachidic\}}) = 0.5134$



Inconsistent with 5 clusters

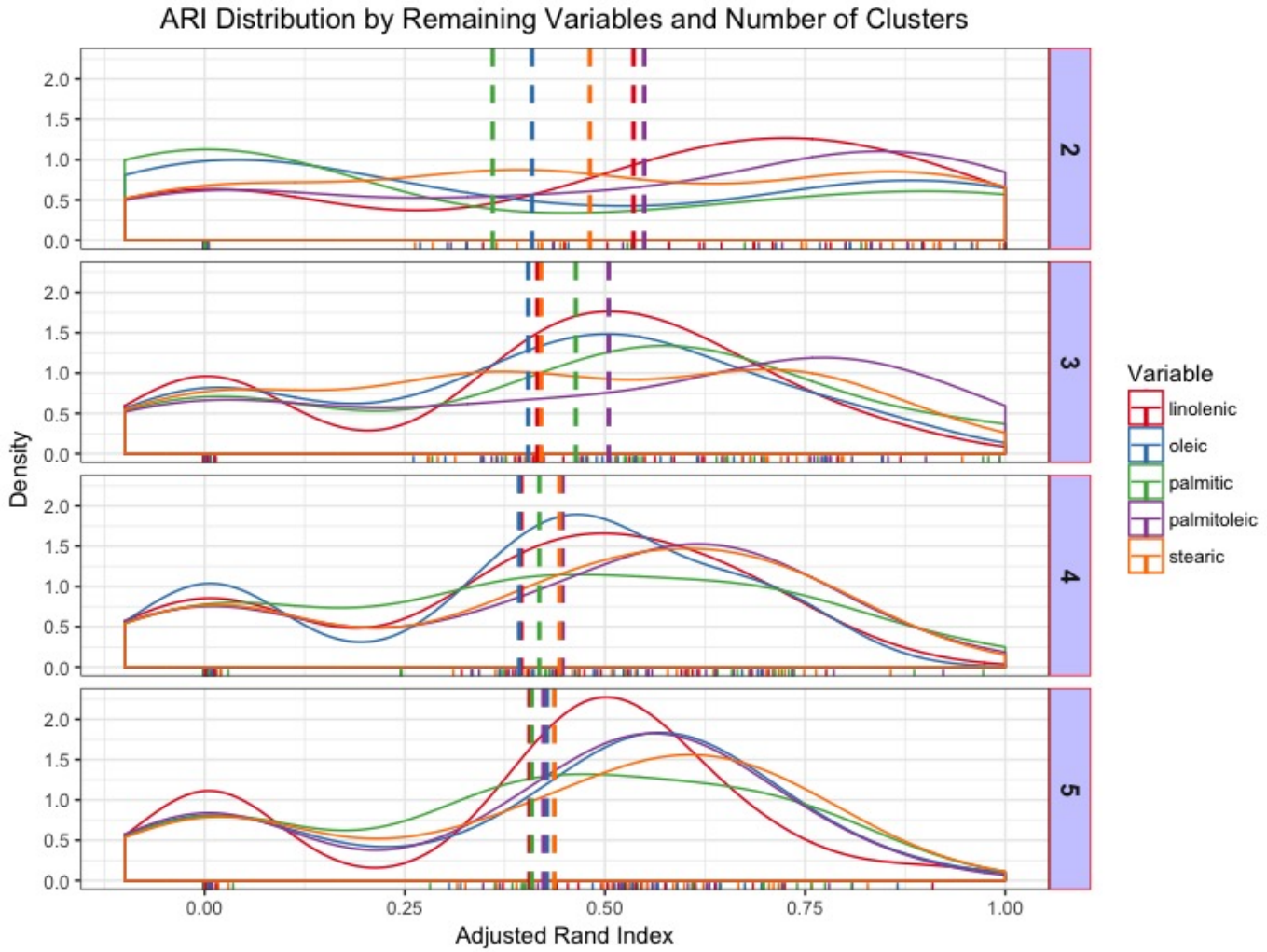
Given {eicosenoic,arachidic}: k=2, linoleic

Stay with 2 clusters, $\overline{ARI}_{2,\{eicosenoic,arachidic,linoleic\}} = 0.5461$



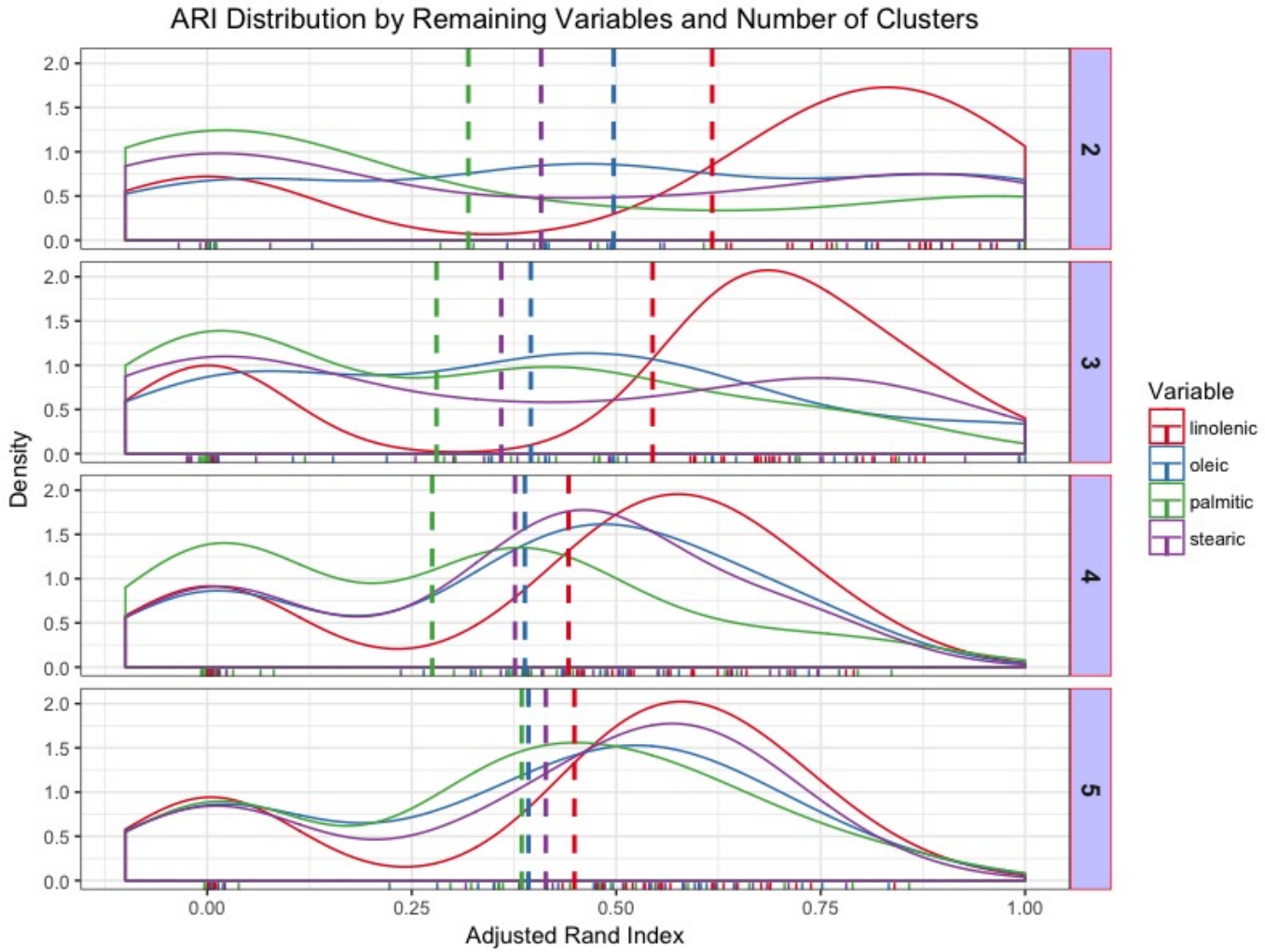
Given {eicosenoic,arachidic,linoleic}: k=2, palmitoleic

Similar results, $\overline{ARI}_{2,\{eicosenoic,arachidic,linoleic,palmitoleic\}} = 0.5491$



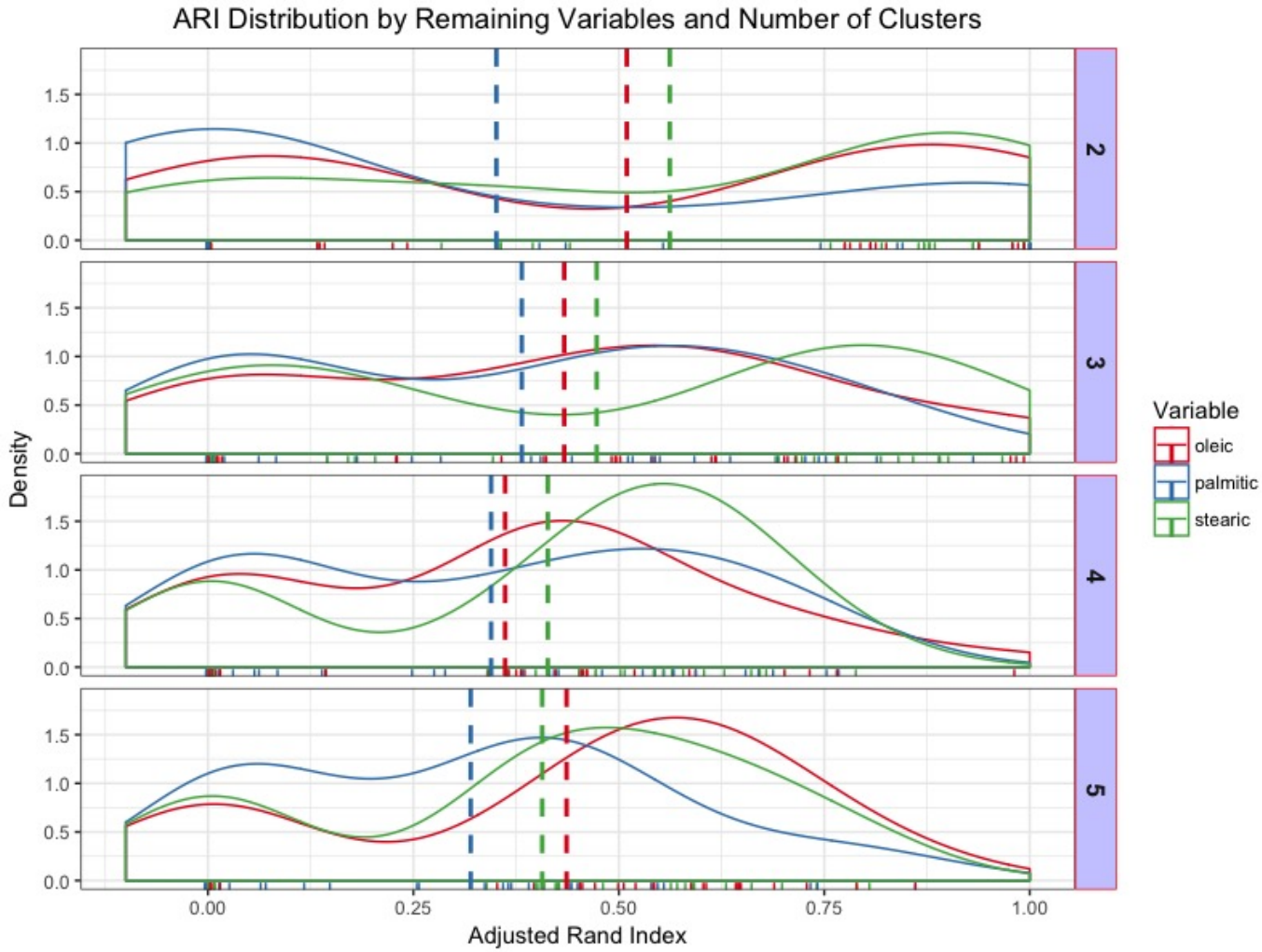
Next: k=2, linolenic

$$\overline{ARI}_{2,\{eicosenoic,arachidic,linoleic,palmitoleic,linolenic\}} = 0.6174$$



Next: k=2, stearic

$$\overline{ARI}_{2,\{eicosenoic,arachidic,linoleic,palmitoleic,linolenic,stearic\}} = 0.5620$$



Stop There

Best option is below 0.5,

$$\overline{ARI}_{5,\{eicosenoic,arachidic,linoleic,palmitoleic,linolenic,stearic,oleic\}} = 0.4733$$