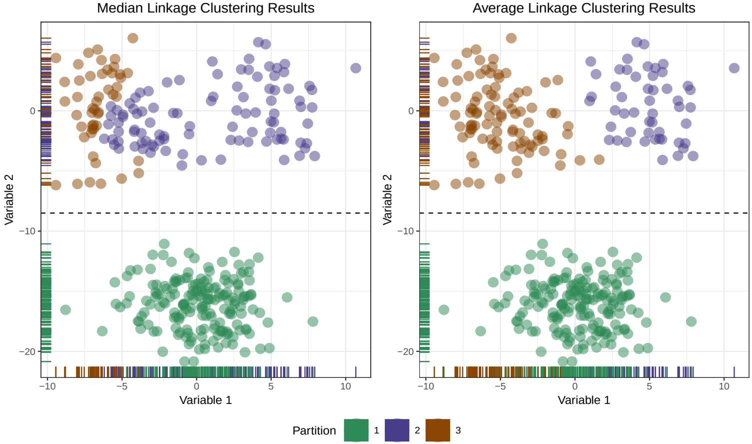
Variable Selection for Consistent Clustering

Ron Yurko Rebecca Nugent Sam Ventura

Department of Statistics & Data Science
Carnegie Mellon University

Symposium on Data Science and Statistics 2018

Variable choice ! inconsistent clusters



Methods disagree using both variables,
but **agree** on two consistent clusters with Variable 2

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

Measuring clustering agreement with ARI

$ARI(p_1; p_2) =$ **Adjusted Rand Index (ARI)**[1], similarity index between two partitions p_1 and p_2

Corrected for chance agreement,

$$\mathbb{E}[ARI(p_1; p_2)] = 0$$

$ARI(p_1; p_2) < 0 \ \rightarrow$ worse than random

$ARI(p_1; p_2) = 1 \ \rightarrow$ identical partitions

[1] [Hubert and Arabie, 1985]

Maximum Clustering Similarity (MCS)[2]

An approach to determine K, number of clusters

Let $M =$ set of clustering methods

Choose K with most frequent max similarity,

$$e.g.\ ARI(p_{1,K}, p_{2,K}) \text{ from } \binom{|M|}{2} \text{ partition pairs}$$

[2] [Albatineh and Niewiadomska-Bugaj, 2011]

Greedy search algorithm for variable selection

Idea: **Greedily search for the most consistent subset of variables across clustering methods and number of clusters** K

Notation:

- $\mathbf{X} = N \times D$ data matrix, $d\ 2\ f1;:::;Dg$
- $S =$ set of selected variables
- $U =$ set of unselected variables, where $S\ [\ U = f1;:::;Dg$ and $S\ \backslash\ U = f;g$
- $M =$ $f$complete, single, Ward, average, McQuitty, median, centroid, kmeans$g$ (*just for illustrative purposes*)

Step 0: Initialize $S = f;g$, $U = f1;:::;Dg$

Greedy search algorithm for variable selection

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

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

Greedy search algorithm for variable selection

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

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

Step 2: Select most consistent result:

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

Greedy search algorithm for variable selection

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

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

Step 2: Select most consistent result:

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

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

Greedy search algorithm for variable selection

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

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

Step 2: Select most consistent result:

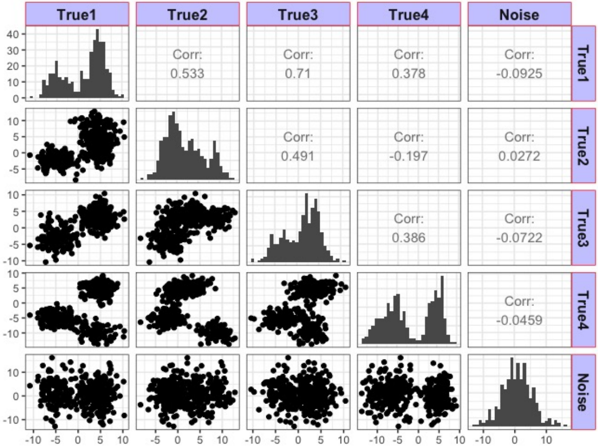
$$d^*, K^* := \underset{d \in U, K}{\arg\max}\ \overline{ARI}_{K, S \cup \{d\}}$$

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

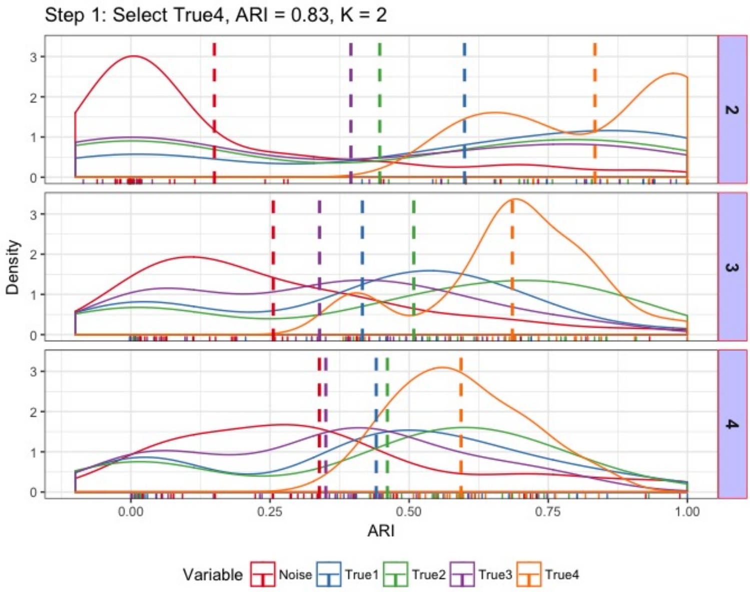
Repeat 1-3 until $U = \{\emptyset\}$ or met stopping criteria

Demo data

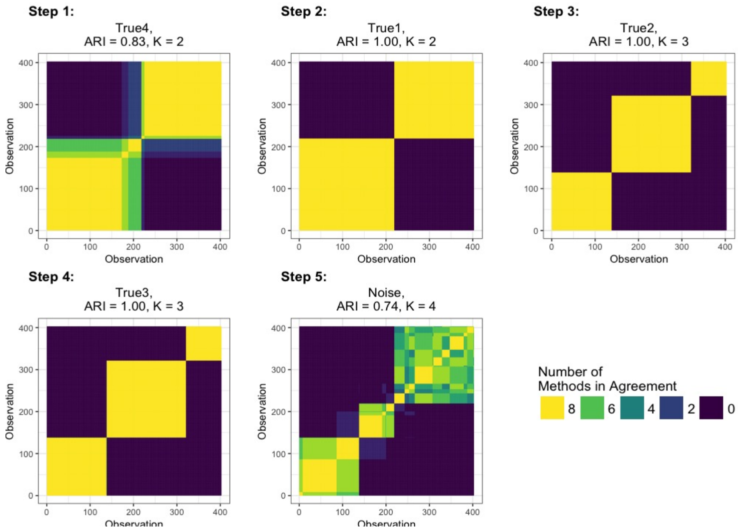
4 true variables, 1 noise variable, and $K = 3$



Step 1 of demo search



Consensus matrices for full search



Bootstrap consistency distributions to address limitations

We want to provide a measure of **con dence** in our decision:

- $f_{K;S}$ = bootstrap distribution for $\overline{ARI}_{K;S}$
- $f_{K;S[fdg}$ = bootstrap distribution for $\overline{ARI}_{K;S[fdg}$
- $overlap(f_{K;S}; f_{K;S[fdg})$ = area of overlap between the two

Bootstrap consistency distributions to address limitations

We want to provide a measure of **con dence** in our decision:

- $f_{K;S}$ = bootstrap distribution for $\overline{ARI}_{K;S}$
- $f_{K;S\,[\,fdg}$ = bootstrap distribution for $\overline{ARI}_{K;S\,[\,fdg}$
- $overlap(f_{K;S}; f_{K;S\,[\,fdg})$ = area of overlap between the two

Include variable d based on distribution **overlap**

IF $9\ d\ 2\ U$ such that $\overline{ARI}_{K;S\,[\,fdg} > \overline{ARI}_{K;S}$ (more consistent)

$$d\ ;K\ := \arg\min_{d2U;K} overlap(f_{K;S}; f_{K;S\,[\,fdg})\ \text{(minimize overlap)}$$

Bootstrap consistency distributions to address limitations

We want to provide a measure of **con dence** in our decision:

- $f_{K;S}$ = bootstrap distribution for $\overline{ARI}_{K;S}$
- $f_{K;S\,[\,fdg}$ = bootstrap distribution for $\overline{ARI}_{K;S\,[\,fdg}$
- $overlap(f_{K;S}; f_{K;S\,[\,fdg})$ = area of overlap between the two

Include variable d based on distribution **overlap**

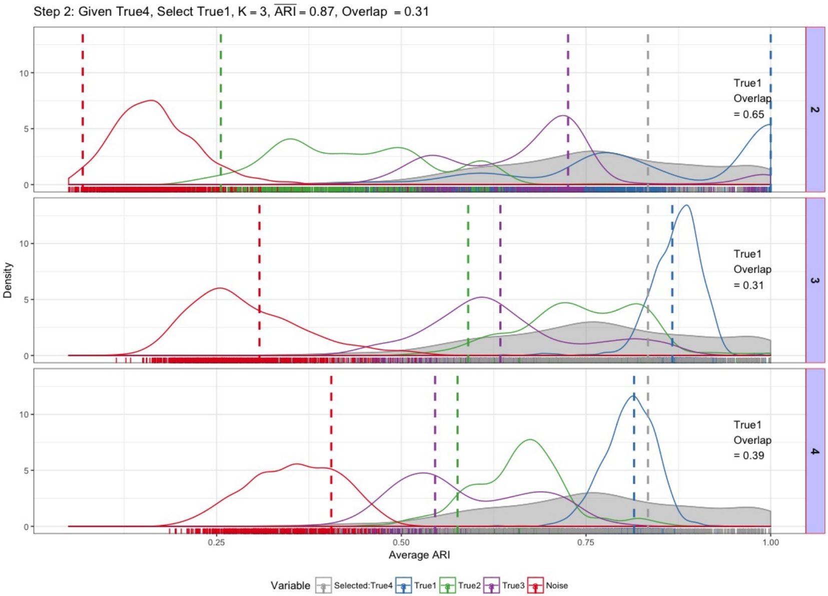
IF $9\ d\ 2\ U$ such that $\overline{ARI}_{K;S\,[\,fdg} > \overline{ARI}_{K;S}$ (more consistent)

$$d\ ;K\ := \underset{d2U;K}{\arg\min}\ overlap(f_{K;S}; f_{K;S\,[\,fdg})\ \text{(minimize overlap)}$$

ELSE (less consistent)

$$d\ ;K\ := \underset{d2U;K}{\arg\max}\ overlap(f_{K;S}; f_{K;S\,[\,fdg})\ \text{(maximize overlap)}$$

Bootstrap distributions for step 2 of demo search



Noise has minimal overlap and is not selected

Step	Variable	$\overline{ARI}$	K	Overlap
1	True4	0.8339	2	-
2	True1	0.8668	3	0.3067
3	True2	1.000	3	0.1338
4	True3	0.9979	3	0.3277
5	Noise	0.7444	4	0.0969

By measuring the overlap, we are confident that including Noise leads to inconsistent clustering results

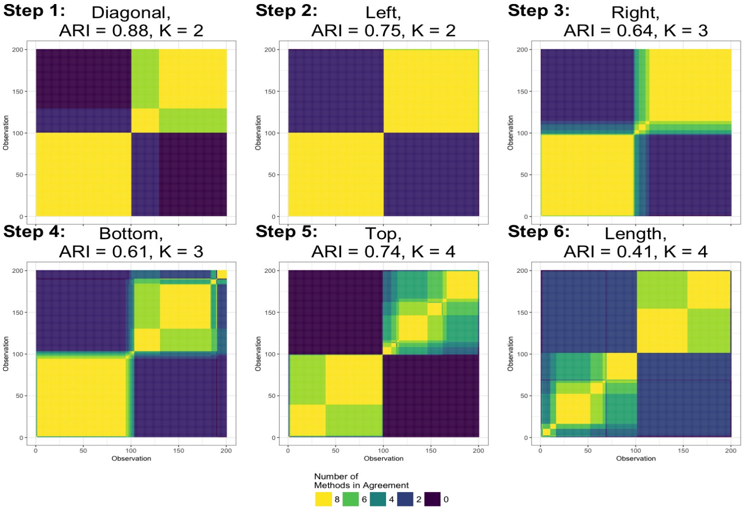
Swiss bank notes example

200 bills that are either counterfeit or real with 6 measurements

Summary of search reveals decrease in consistency:

Step	Variable	$\overline{ARI}$	K	Overlap
1	Diagonal	0.8755	2	-
2	Left	0.7500	2	0.8969
3	Right	0.6418	2	0.8789
4	Bottom	0.6112	3	0.6401
5	Top	0.7438	4	0.7262
6	Length	0.4113	4	0.7916

Swiss bank notes consensus matrices



Future Work

Simulation study, examine properties of ARI values

Explore different notions of stopping criteria

- Only considered average, but distributions are multimodal and assymetrical (e.g. mass above threshold?)

Inclusion of removal step

Consider sensitivity to different types of clustering methods

- What about soft partitions?[3]

[3] [Flynt et al.,]

Thanks and References I

Albatineh, A. N. and Niewiadomska-Bugaj, M. (2011).
Mcs: A method for nding the number of clusters.
Journal of Classi cation, 28:184{209.

Andrews, J. L. and McNicholas, P. D. (2014).
Variable selection for clustering and classi cation.
Journal of Classi cation, 31(2):136{153.

Efron, B. (1979).
Bootstrap methods: Another look at the jackknife.
The Annals of Statistics, 7(1):1{26.

Flynt, A., Dean, N., and Nugent, R.
sari: An agreement measure for pairs of class assignments incorporating posterior probabilities.

Thanks and References II

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.

Hahsler, M., Buchta, C., and Hornik, K. (2017).
seriation: Infrastructure for Ordering Objects Using Seriation.
R package version 1.2-2.

Hubert, L. and Arabie, P. (1985).
Comparing partitions.
Journal of Classi cation, 2:193{218.

Hurley, C. (2012).
gclus: Clustering Graphics.
R package version 1.3.1.

Thanks and References III

Milligan, G. W. (1996).
Clustering validation: Results and implications for applied analyses.
In Arabie, P., Hubert, L., and Soete, G. D., editors, *Clustering and Classi cation*, pages 341{375. World Scienti c, Singapore.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003).
Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data.
Journal of Machine Learning, 52:91{118.

Morey, L. C. and Agresti, A. (1984).
The measurement of classi cation agreement: An adjustment to the rand statistic for chance agreement.
Educational and Psychological Measurement, 44(1):33{37.

Pastore, M. (2017).
overlapping: Estimation of Overlapping in Empirical Distributions.
R package version 1.5.0.

Thanks and References IV

Qiu, W. and Joe, H. (2006).
Separation index and partial membership for clustering.
Computational Statistics and Data Analysis, 50:585{603.

Qiu, W. and Joe, H. (2015).
clusterGeneration: Random Cluster Generation (with Speci ed Degree of Separation).
R package version 1.3.4.

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

Witten, D. M. and Tibshirani, R. (2010).
A framework for feature selection in clustering.
Journal of the American Statistical Association, 105(490):713{726.