Choosing the Number of Clusters in Monothetic Clustering

Tan Tran & Mark Greenwood
Department of Mathematical Sciences, Montana State University

Clustering

- Clustering algorithms attempt to group subjects so that the dissimilarity within clusters is smallest while the between cluster dissimilarities are largest.
  - Q = number of response variables, n = sample size, y_{iq} is the i^{th} observation on variable q.
  - n objects in Ω, with partition $P_J = C_1, \ldots, C_J$ the J-cluster partition of Ω.
- The choice of J impacts the group memberships and thus the interpretation of the results.
  - J too small puts "unlike" subjects together.
  - J too large splits observations that should be together.
- Ruspini (1970) used a data set with 70 observations on two variables to demonstrate clustering techniques.

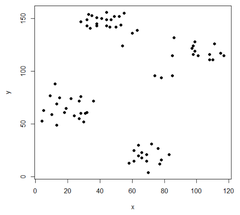


Figure: Ruspini data set

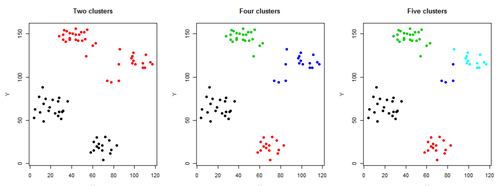


Figure: Different possible cluster solutions for Ruspini's data.

Research Question

How many clusters should observations be classified into when using monothetic clustering?

- Goal: Objective numerical measures for selecting the "correct" number of clusters.

Monothetic Cluster Analysis (Chavent, 1998)

- Hierarchical, recursive partitioning of multivariate responses based on binary decision rules built from individual response variables.
- Inspired by Regression Trees (see Hastie et al., 2009).
  - Search for splits from each response variable that provide the best split of the multivariate responses (measured with global criterion called inertia).
  - Recursively apply to each sub-partition, recording splitting rules that define clusters.
  - Generates a set of binary rules for cluster membership ("monothetic").
- Defining within cluster inertia: total variability around the cluster mean.
  - If Euclidean distance is being used, inertia for cluster j,

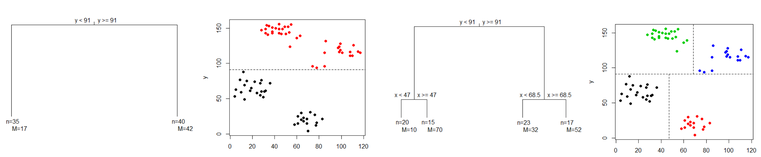
$$I(C_j) = \sum_{i \in \text{Cluster}_j} \sum_{q=1}^{Q} (y_{iq} - \overline{y}_{\cdot q})^2.$$

  - Choose the bipartition for cluster C to maximize the difference in inertia,

$$I(C) - I(C_1) - I(C_2).$$

  - Recursively apply to each sub-partition.
  - Equivalence of distances for all observations within cluster and within group variation (p. 388, James et al., 2013) is exploited to use inertia with any dissimilarity matrix.
- An R package `MonoClust` is under development - go to `http://tinyurl.com/MonoClust`.

Monothetic Cluster Analysis of Ruspini Data



(a) One split, two clusters (b) Three splits, four clusters

Figure: Binary partitioning tree with maps of optimal splits.

- For four clusters:
  - Cluster 1: y below 91 and x below 47 and Cluster 2: y below 91 and x above 47
  - Cluster 3: y over 91 and x below 68.5 and Cluster 4: y over 91 and x over 68.5

Some Popular Metrics for Choosing Number of Clusters

1. Average silhouette width (AW) (Rousseeuw, 1987)
   - The silhouette width is a measure of how "comfortable" an observation is in the cluster it resides in.
     - Between 0 and 1 if "happiest" in existing cluster.
     - 0 if ambivalent about cluster membership vs. next closest cluster.
     - Between -1 and 0 if observation "wants to leave" current cluster.
   - The average silhouette width is mean of all n silhouette widths.
   - Select cluster solution size J that has maximum average silhouette width.
   - It is not defined for $J = 1$ so cannot select a single cluster solution.
2. Calinski and Harabasz (CH)'s pseudo-F (Calinski and Harabasz, 1974)
   - Choose J to maximize variation between clusters relative to variation within clusters.
   - Use pseudo-F= $\frac{B(J)/(J-1)}{W(J)/(n-J)}$ with $B(J)$ the between cluster sums of squares (possibly from dissimilarities) and $W(J)$ the within cluster sums of squares.
   - It is not defined for $J = 1$ so cannot select a single cluster solution.
3. Many others exist: See Milligan and Cooper (1985) and `NbClust` (Charrad et al., 2014) in R.

Proposed Methods for Choosing Number of Clusters

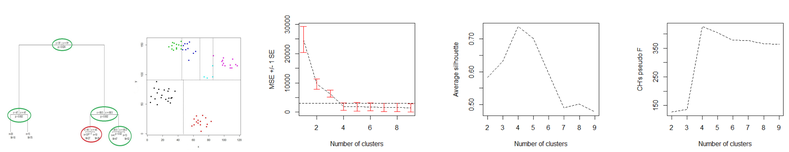
1. M-fold Cross-Validation
   - Based on ideas for pruning regression trees (Breiman et al., 1984).
   - Randomly divide data set into M equal-sized subsets, withhold a subset, and use the rest for training the method.
   - Compute a measure of prediction error for the m^{th} set of withheld observations, such as $MSE_m = \frac{1}{n_m} \sum_{q=1}^{Q} \sum_{i \in m} (y_{iq} - \hat{y}_{iq})^2$ where $\hat{y}_{iq}$ are the predicted responses.
   - Cross-validation based estimate of the error for the J cluster solution is

$$CV_J = \frac{1}{M} \sum_{m=1}^{M} MSE_m.$$

   - CV-based selection rules:
     1. Choose J that provides the smallest CV_J (`minimum CV rule`).
     2. Choose solution that provides simplest solution within 1 standard error of the minimum (`1 SE rule`) where $SE = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (MSE_i - \overline{MSE})^2}$.
   - Monothetic clustering is one of the few clustering algorithms that provides a clear prediction rule – use the binary rules to assign a new observation to a cluster.
     - Then use cluster means to generate predicted response $\hat{y}_{iq}$.
     - Implemented as `predict` for MonoClust class (`predict.MonoClust`).
   - Open question: How does the maximum size of the tree explored impact 1SE rule results?
2. Hypothesis tests at each bipartition in the monothetic cluster solution
   - Based on ideas from conditional inference trees (R package `party`, Hothorn et al., 2006).
     - Grow tree until a split has a p-value over pre-determined threshold (say α of 0.01 or 0.05), adjusting p-values based on depth in tree.
   - Test assesses H_0: *no difference in two groups at node* using permutations of observations and pseudo-F test statistic (as defined in perMANOVA in Anderson, 2001).
     - Available in `adonis` in `vegan` package (Oksanen et. al., 2015).
   - P-values are Bonferroni adjusted for number of tests required to get to level of tree,

$$\text{P-value}_{adj} = \text{depth} \times \frac{\text{No. of } (F_{perm} \geq F_{observed})}{\text{No. of perm}}.$$

   - Problem with applying regular tests to monothetic clustering:
     - Selected split is always the most extreme result possible on variable that defined the bipartition.
     - Potential solution: Only use variation from the Q-1 variables not used to define two groups in test statistic.
       - Differences will be detected if the binary split is useful on other variables.
       - Can not be applied to clustering when Q=1.
   - Implemented as `perm.test` which adds p-value information to the tree-display with `auto.pick` argument to automatically prune the tree based on a selected threshold.



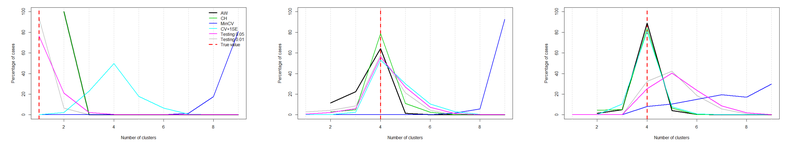
(a) Hypothesis testing: selects 5 clusters (b) 10-fold CV: selects max (minCV), 4 (1SE) (c) Ave silhouette: selects 4 clusters (d) CH's pseudo-F: selects 4 clusters

Figure: Criteria results for monothetic cluster analysis of Ruspini's data.

Simulation Study: Three scenarios considered, L^2 distances

1. *Null model with 10 dimensions:* J=1, Q=10, $y_{iq} \sim U(-1, 1)$
2. *Four clusters in 3 dimensions:* $J = 4$, Q=3, $\mu_j \sim N(0, 5\text{I}_3)$ and $\mathbf{y}_i \sim N(\mu_j, \text{I}_3)$.
   - Cluster sizes are randomly chosen from either 25 or 50 observations.
   - Only samples where distance between two closest observations in clusters is at least 2 units are analyzed.
3. *Four clusters in 10 dimensions:* $J = 4$, Q=10, $\mu_j \sim N(0, 1.9\text{I}_{10})$ and $\mathbf{y}_i \sim N(\mu_j, \text{I}_{10})$
   - Cluster sizes randomly chosen from 25 or 50 observations and distance between clusters ≥ 2 units.

Simulation Study: Results of 500 simulations



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure: Simulation results

- Only hypothesis testing method works well when $J = 1$.
  - It is quite successful when $\alpha = 0.01$ at correctly preventing a search past the root node.
  - With $\alpha = 0.01$, the $J = 1$ solution is rejected 6% of the time – type I error rates inflated because assessing splits that are optimal globally from given choices.
    - Rejection rates were much higher with moderate Q when all variables were included in test statistic.
- Hypothesis testing using the more conservative α value provided more correct results but hypothesis testing struggled with larger Q.
- The minimum CV value is often the largest number of clusters considered – suggests criterion does not provide much penalty for over-fitting.
- 1SE rule with CV is much better than minimum CV and close to other methods in performance except in *Scenario 1*.
  - Some mistakes in *Scenarios 2* and *3* might be due to var(CV) in different random partitions.

Conclusions

- Some methods, such as CH and AW, cannot select one cluster so can't help researcher decide between one versus more than one cluster.
  - Clustering algorithms always generate groups – and these methods can't assess whether any should be found.
- Both CV and p-value based methods are attractive as they provide direct information about utility of single cluster versus other sizes.
  - Hypothesis testing can work but seems to suffer as the dimension of the data set grows.
  - CV results vary in repeated application with same data; variability may be impacting selection performance.
- The speed of `MonoClust` depends largely on the number of dimensions.
- CV is slowest of methods considered because of M re-fittings of monothetic cluster solution.
- Permutation testing is fast due to `adonis` function and only single monothetic cluster solution required.
- Recommendations for finding number of clusters in monothetic clustering:
  - Use permutation test to assess evidence for more than one cluster using modified test statistic and small α.
  - If initial p-value is small, use pseudo-F maximization (CH) to find correct size of cluster solution.
  - Use predict function to classify and generate predictions for new observations.
  - If using CV-based methods, use 1SE rule and consider multiple runs of CV splitting process to identify the most commonly selected optimal cluster size.

Future Work

- Finalize the R package and assess criteria in more situations with changing variability.
- Include GAP (Tibshirani et al., 2001) and other criteria in comparisons - some may work better for monothetic clustering than others and provide $J = 1$ selection possibility.
- Permutation tests could be used in other hierarchical clustering techniques, assessing evidence of need for divisions from top of tree down.
  - See `pvclust` (Suzuki et al., 2006) for bootstrap-based p-value approach with regular hierarchical clustering.

Acknowledgments

Sabbatical support for Greenwood from Montana State University in Fall 2014 and travel support from the College of Letters and Sciences and Department of Mathematical Sciences.

References, copy of poster, and MonoClust code available in github repository: `http://tinyurl.com/MonoClust`.