Estimating K with the Gap Statistic

Ryan Hicks

Department of Statistics, Colorado State University

December 3, 2015

Motivation

- Minimizing within-cluster variation does not work

- $W(K) = \sum_{k=1}^{K} \sum_{i \in C_k} \parallel X_i - \overline{X_k} \parallel_2^2$

- Maximizing between-cluster variation does not work

- $B(K) = \sum_{k=1}^{K} |C_k| || \overline{X}_k - \overline{X} ||_2^2$

Calinski-Harabasz Index

- $CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}$

- Choose a maximum number of clusters then find $\hat{K} = \arg\max_{K \in \{2,\ldots,K_{max}\}} CH(K)$

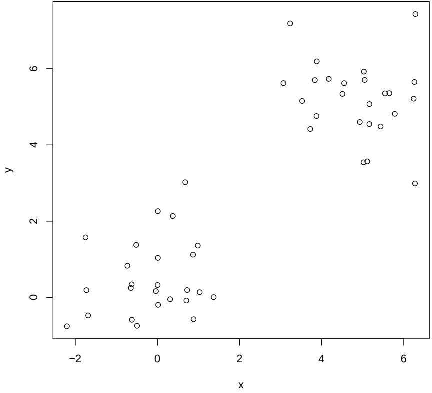
- However, CH(K) is undefined for K = 1; a big disadvantage.

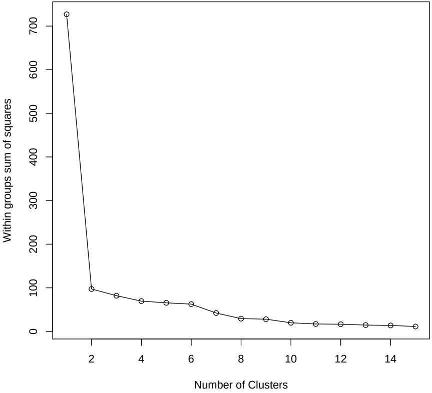
- There may not be any underlying clusters in the data.

Gap Statistic

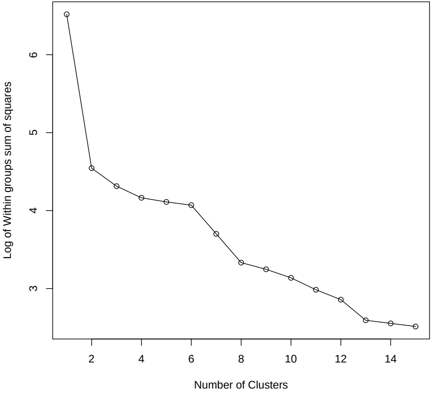
- $\forall i, 1 \leq i \leq K_{max}$, run a clustering method on the dataset to find i clusters, and sum the distance of all points from their cluster mean.
- Generate B reference datasets, easily found by uniformly sampling from a bounding rectangle of the original dataset, though there are more complex approaches.
- Define the gap statistic by $Gap_n(k) = E_n^*\{ln(W_k)\} - ln(W_k)$
- Choose the number of clusters that maximizes the gap.

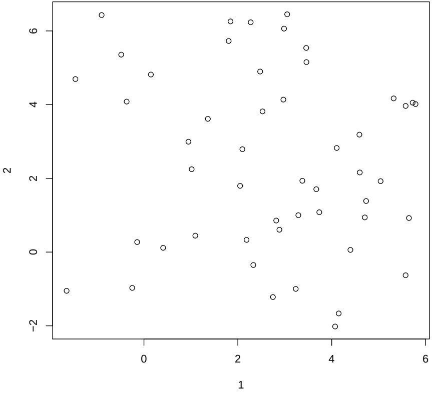
Example (Using K-means)



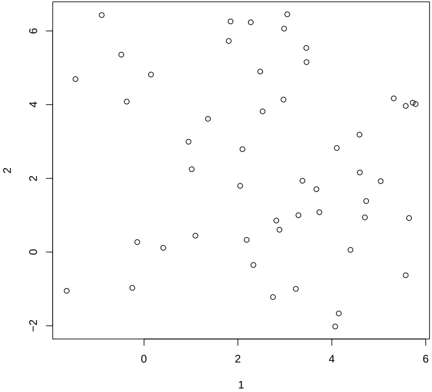


Example

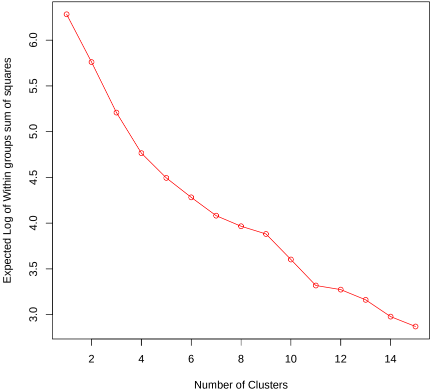

Log of Within groups sum of squares
6
5
4
3
2
4
6
8
10
12
14
Number of Clusters


2
6
4
2
0
−2
0
2
4
6
1

Example

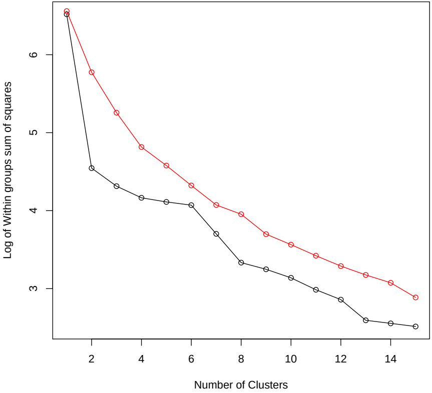


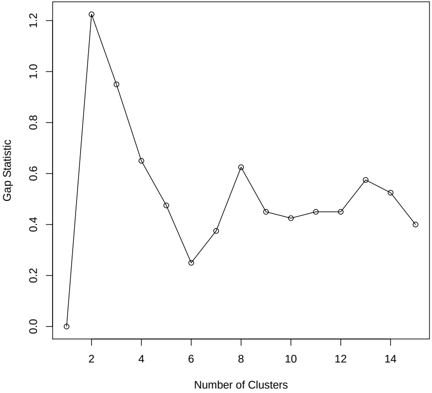
6
4
2
0
−2
2
0
2
4
6
1



6.0
5.5
5.0
4.5
4.0
3.5
3.0
Expected Log of Within groups sum of squares
2
4
6
8
10
12
14
Number of Clusters

Example


Log of Within groups sum of squares
3
4
5
6
2
4
6
8
10
12
14
Number of Clusters


Gap Statistic
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2
4
6
8
10
12
14
Number of Clusters

The End

Thank You!

Questions?