Estimating the number of clusters in a data set via the gap statistics

Rudraksha Samdhani*
Rudraksha.Samdhani@campus.lmu.de
Ludwig Maximilian University
Munich, Bayern, Germany

ABSTRACT

A method (gap statistics) to estimate the optimal number of clusters in any data set. The method uses output of any clustering algorithm and compares the change in within cluster dispersion with that expected under a appropriate reference distribution.

KEYWORDS

Clustering, Groups, Hierarchy, K-Means, Uniform Distribution

ACM Reference Format:
Rudraksha Samdhani. 2018. Estimating the number of clusters in a data set via the gap statistics. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Cluster analysis is an important tool for "unsupervised" learning. The problem of finding groups in data without the help of response variable.

A major challenge here is the estimation of optimal number of clusters. Fig.1 here shows a typical gap plot of a 2 cluster data set. We can see the gap is maximum at $k = 2$ and then keeps on decreasing or does not exceed the maximum.

This paper proposes the 'gap' method for estimating the number of clusters. It is applicable on any clustering algorithm but for simplicity the theoretical part of this analysis will focus on K-means.

2 THE GAP STATISTIC

The gap statistic compares the total intracluster variation for different values of k with their expected values under null reference distribution of the data (i.e. a distribution with no obvious clustering).

For simplicity of the theoretical part of the analysis we will be using the example from "**Estimating the number of clusters in a data set via gap statistic**, Robert Tibshirani, Guenther Walther, Trevor Hastie" but more examples with code will be mentioned in the examples section.

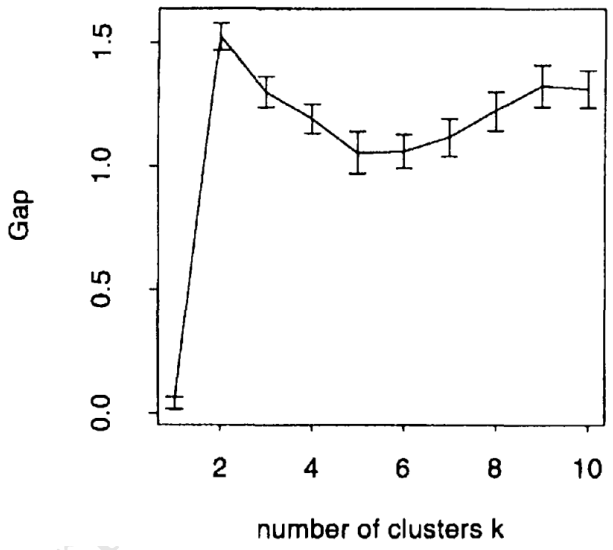

Figure 1: Two-Cluster example

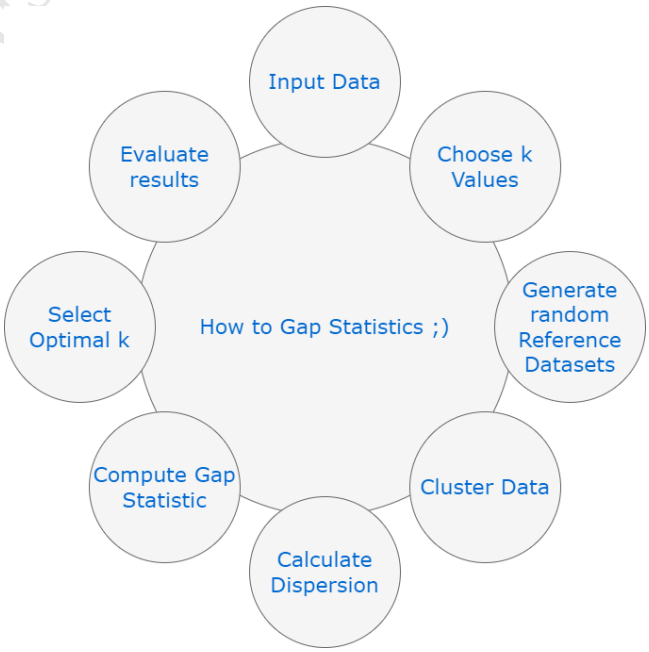

Figure 2: Gap Statistics in a nutshell

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2.1 Theoretical and Mathematical interpretation of Gap Statistics

The data consists of p features measured on n different observations i and i'. We are taking the squared eucledian distance for $d_{ii'}$.
Lets say we cluster the data into k clusters $C_1, ..., C_k$ with C_r denoting the indices of observations in cluster r, and $n_r = |C_r|$.
The sum of pairwise distances for all points in cluster r is denoted by

$$D_r = \Sigma_{ii'\epsilon C_r} d_{ii'}$$

,and set

$$W_k = \Sigma_{r=1}^{k} \frac{1}{2n_r} D_r$$

2.2 The basic idea behind this approach

The idea of this approach is to standardize the graph of $log(W_k)$ by comparing it with its expectation under an appropriate null reference distribution of data. The estimate of the optimal numbers of clusters is then the value of k for which $log(W_k)$ falls the farthest below this reference curve. Hence

$$Gap_n(k) = E_n^* log(W_k) - log(W_k)$$

where E_n^* denotes expectation under a sample size of n from the reference distribution. The estimate $\hat{k}$ will be the value maximizing the $Gap_n(k)$.
If the data actually have K well separated clusters, we expect $log(W_k)$ to decrease faster than its expected rate for $k <= K$. With $k > K$ there's one more cluster added and $log(W_k)$ should decrease *more slowly*. Hence the gap statistic will be largest at $k = K$.
To operate gap statistic we need to find an appropriate reference distribution and to assess the sampling distribution of the gap statistic.

3 THE REFERENCE DISTRIBUTION

The reference distribution in the gap statistic is a distribution of within-cluster dispersions that we would expect to see by random chance in a dataset without any apparent clustering structure.

3.1 How to find appropriate reference distribution

The population version of gap statistic in the case of K-Means clustering

$$g(k) = \frac{log[MSE_{X^*}(k)]}{log[MSE_{X^*}(1)]} - \frac{log[MSE_X(k)]}{log[MSE_X(1)]}$$

The logarithms of the variances have been subtracted to make $g(1) = 0$. So we are looking for a least favorable single-component reference distribution on X^* such that $g(k) <= 0$. The two theorems coming show the possibilities of reference distributions.

3.2 Theorem 1.

Let $p = 1$. Then for all $k >= 1$

$$inf_{X\epsilon S^p}[\frac{log[MSE_X(k)]}{log[MSE_X(1)]}] = \frac{log[MSE_U(k)]}{log[MSE_U(1)]}$$

In other words, among all unimodal distributions, the uniform distribution is most likely to produce spurious clusters by gap test.

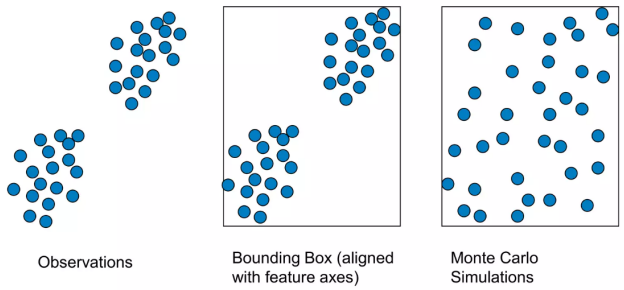

Figure 3: Types of Simulations

3.3 Theorem 2.

If $p > 1$ then no distribution $U\epsilon S^p$ can satisfy the infimum equation unless its support is degenerate to a subset of a line.
The assertion of the last theorem is not contingent on our definition S^p of a single-component model. Simple calculations show that employing a reference distribution with degenerate support will result in an ineffectual procedure.

3.4 Types of Simulations

1. Bounding Box : A BB is a rectangle that encloses an object. Intuitively, a ground-truth BB can be represented as [x,y,w,h] where x, y are coordinate points of the BBs, w and h are the width of the BB. Coordinates of the BBs corners are calculated with reference to the upper-left corner of the image with coordinates (0,0).
2. Monte-Carlo : A Monte Carlo simulation is used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. It is a technique used to understand the impact of risk and uncertainty.
Fig. 3 shows both of the types of simulations. **We will be using Monte Carlo Simulations in our calculations and computations.**

4 COMPUTATIONAL IMPLEMENTATION OF GAP STATISTIC

The computation of gap statistic is as follows :
Step 1 : cluster the observed data, varying the total number of clusters from $k = 1, 2, ..., K$, giving within-dispersion measures $W_k, k = 1, 2, ..., K$.
Step 2 : generate B reference data sets, using the uniform prescription from (a) or (b) and cluster each one giving within-dispersion measures $W_{kb}^*, b = 1, 2, ..., B, k = 1, 2, ...K$.
Compute the estimated gap statistic :

$$Gap(k) = (1/B)\Sigma b log(W_{kb}^*) - log(W_k)$$

Step 3 : let $\bar{l} = (1/B)\Sigma b log(W_{kb}^*)$, compute the standard deviation :

$$sd_k = [(1/B)\Sigma b(log(W_{kb}^*) - \bar{l})^2]^{1/2}$$

and define $s_k = sd_k\sqrt{(1 + (1/B))}$. Finally choose the number of clusters via :

$$\hat{k} = \text{smallest } k \text{ such that } Gap(k) >= Gap(k+1) - s_k + 1$$

Fig.4 shows a small flowchart of this process.

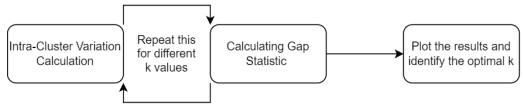

Figure 4: Computation of Gap statistic

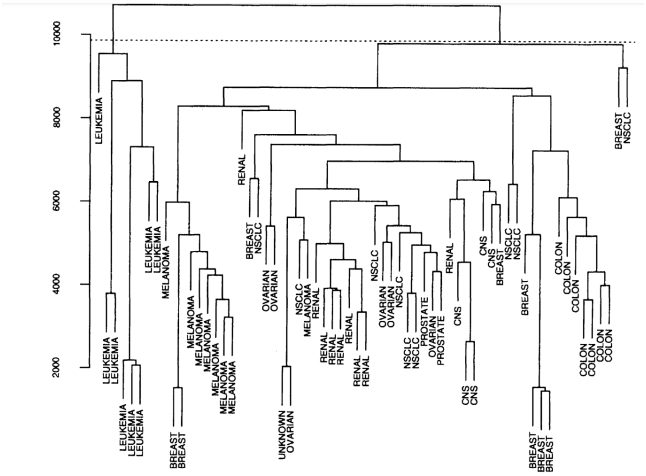

Figure 5: Dendogram from the DNA microarray data. The dotted line cuts the tree, leaving two clusters as suggested by gap statistic.

5 EXAMPLES

5.1 Application to hierarchical clustering and DNA Microarray data

This example is from "**Estimating the number of clusters in a data set via gap statistic**, Robert Tibshirani, Guenther Walther, Trevor Hastie". The data is 6834 x 64 matrix of gene expression measurements. The data is taken from Ross et al. (2000). By applying hierarchical clustering to the coloums, using squared error and average linkage, we obtain the dendogram in Fig.5. Not surprisingly many clusters of the same type are clustered together.

The results of gap statistic are shown in Fig.6, . The estimated number of clusters is 2. However the gap function rises again after 6 clusters, suggesting there are 2 well separated clusters and more less separated ones.

Fig.7 shows how observed and expected $log(W_k)$ changes wrt. K.

5.2 USArrests Data set

In this example we apply K-Means algorithm and gap statistic on the US Arrests data set available on R directly. We applied K Means algorithm on the original and the reference datasets and calculated the gap for different K values. The results are of this are in Fig.8.

Fig.9 shows the final clusters that are formed with K-Means as the clustering algorithm taking in account the 4 optimal clusters suggested by gap statistic.

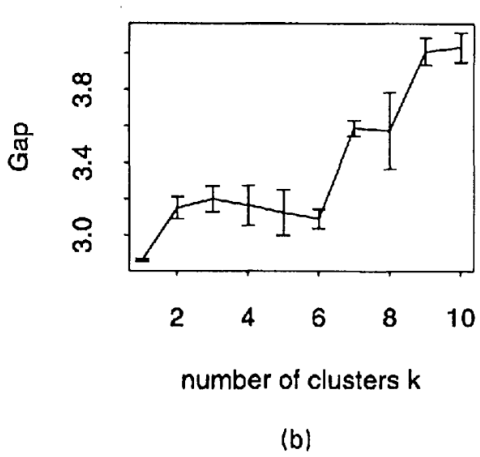

Figure 6: The Gap Statistic for DNA microarray

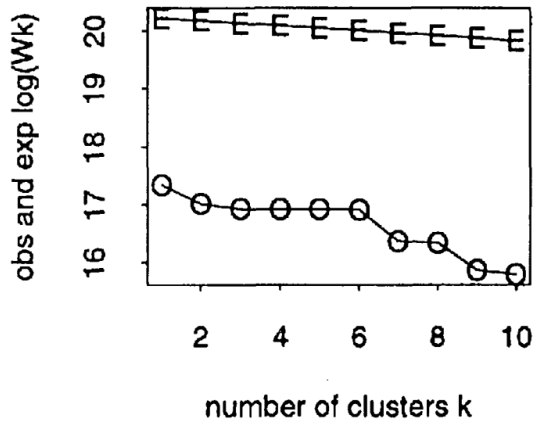

Figure 7: expected and observed $log(W_k)$ of DNA microarray

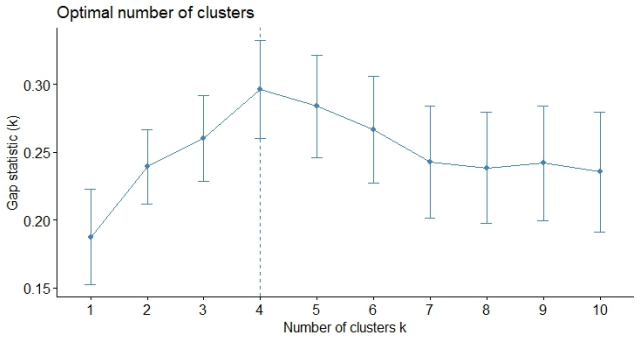

Figure 8: The Gap Statistic for USArrests Data set. Optimal number of clusters 4.

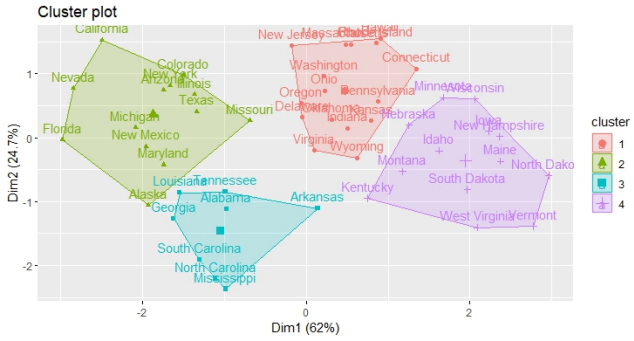

Figure 9: The 4 Clusters in the data set US Arrests.

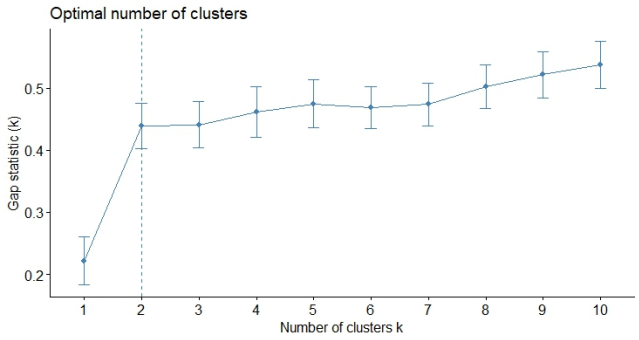

Figure 10: the gap statitic for the Rock Example (two cluster data set)

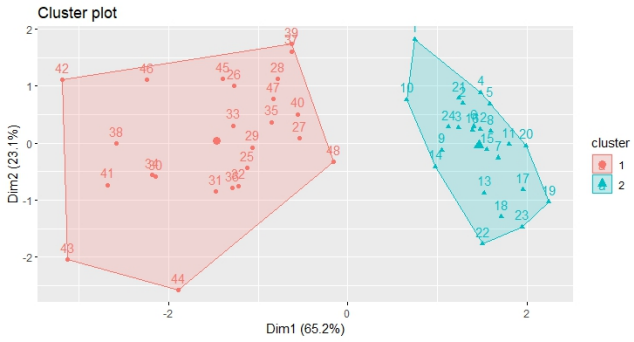

Figure 11: the two final clusters of Rock example

5.3 Two cluster and No cluster

I have also simulated a two cluster and a no cluster data set with K-Means as clustering algorithm and computed the gap statistic for the data. For a two cluster data set we find that the gap curve has a clear maximum at $\hat{k} = 2$. Fig.10 shows the gap plot and Fig.11 shows the final two clusters of the data set. The data used here is the rock data set available in R library.

For the no cluster or 1 whole cluster or unclustered data set we see that the observed and expected curves are very close and the $\hat{k} = 1$. The gap keeps on decreasing and is negative wrt. K. Note that here the data is synthetic for the example. Fig.12 shows the expected and observed $log(W_k)$ and Fig.13 shows the negative gap curve.

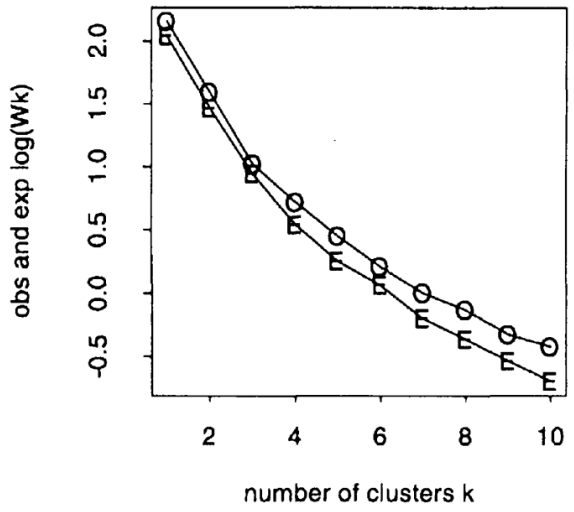

Figure 12: observed and expected $log(W_k)$ for no cluster data

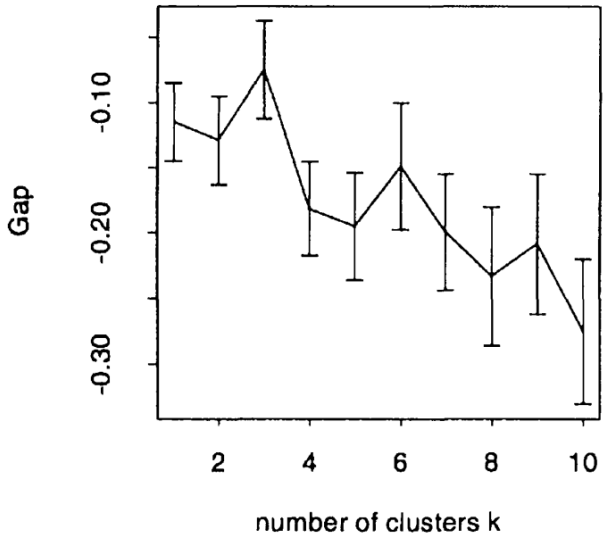

Figure 13: The negative gap curve for the no cluster data

6 OTHER APPROACHES

There are a lot of other methods proposed to estimate the number of clusters in a data set. Most global methods have the disadvantage that they are undefined for one cluster and hence offer no indication whether the data should be clustered at all.

6.1 Milligan and Cooper (1985)

They carried out a comprehensive simulation comparison of 30 different procedures. The best performance was given by Calinski and Harabasz(1974) index :

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

where B(k) and W(k) are between and within cluster sum of squares, with k clusters.

6.2 Krzanowski and Lai(1985)

They proposed another criterion for choosing the number of clusters. This followed a proposal by Marriot(1971), who used the determinant rather than the trace, of the within sum of squares matrix.

$$DIFF(k) = (k-1)^{2/p}W_{k-1} - k^{2/p}W_k$$

and chose k to maximize the quantity :

$$KL(k) = |\frac{DIFF(k)}{DIFF(k+1)}|$$

6.3 Hartigan(1975)

They also proposed another statistic to estimate the optimal number of clusters in a data set.

$$H(k) = [\frac{W(k)}{W(k+1)} - 1]/(n-k-1)$$

6.4 Kaufman and Rousseeuw(1990)

They proposed the silhouette statistic, for accessing the clusters and estimating the optimal number. For obv. i, let $a(i)$ be the average distance to other points in the cluster, and $b(i)$ the average distance to the points in the nearest cluster. Then the silhouette statistic is defined as :

$$s(i) = \frac{b(i)-a(i)}{max[a(i),b(i)]}$$

The value maximaizing the average is considerd to be the optimal number of clusters.

7 SIMULATIONS

In the paper "**Estimating the number of clusters in a data set via gap statistic**, Robert Tibshirani, Guenther Walther, Trevor Hastie" they generated the data sets into 5 different senarios :

- *null data in 10 dimensions* – 200 data points uniformly distributed over the unit square in 10 dimensions.
- *three clusters in 2 dimensions* – the clusters are standard normal variables with (25,25,50) observations, centered at (0,0), (0,5) and (5,-3).
- *four clusters in 3 dimensions* – each cluster was randomly chosen to have 25 or 50 standard normal observations, with centers randomly chosen as $N(0, 5I)$.
- *four clusters in 10 dimensions* – each cluster was randomly chosen to have 25 or 50 standard normal observations, with centers randomly chosen as $N(0, 1.9I)$.
- *two elongated clusters in three dimensions* – each cluster specifically generated. Exact definition in paper.

They generated a table of results with all these simulations for different methods including Gap/unif with uniform reference distribution over range of each observed feature and Gap/pc using uniform reference in principal components orientation.
Fig.14 shows the table from the paper "**Estimating the number of clusters in a data set via gap statistic**, Robert Tibshirani, Guenther Walther, Trevor Hastie" with different methods and different simulations

Method	*Estimates of the following numbers of clusters k̂:*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Null model in 10 dimensions										
CH	0‡	50	0	0	0	0	0	0	0	0
KL	0‡	29	5	3	3	2	2	0	0	0
Hartigan	0‡	0	1	20	21	6	0	0	0	0
Silhouette	0‡	49	1	0	0	0	0	0	0	0
Gap/unif	49‡	1	0	0	0	0	0	0	0	0
Gap/pc	50‡	0	0	0	0	0	0	0	0	0
3-cluster model										
CH	0	0	50‡	0	0	0	0	0	0	0
KL	0	0	39‡	0	5	1	1	2	0	0
Hartigan	0	0	1‡	8	19	13	3	3	2	1
Silhouette	0	0	50‡	0	0	0	0	0	0	0
Gap/unif	1	0	49‡	0	0	0	0	0	0	0
Gap/pc	2	0	48‡	0	0	0	0	0	0	0
Random 4-cluster model in 3 dimensions										
CH	0	0	0	42‡	8	0	0	0	0	0
KL	0	0	0	35‡	5	3	3	3	0	0
Hartigan	0	1	7	3‡	9	12	8	2	3	5
Silhouette	0	20	15	15‡	0	0	0	0	0	0
Gap/unif	0	1	2	47‡	0	0	0	0	0	0
Gap/pc	2	2	4	42‡	0	0	0	0	0	0
Random 4-cluster model in 10 dimensions										
CH	0	1	4	44‡	1	0	0	0	0	0
KL	0	0	0	45‡	3	1	1	0	0	0
Hartigan	0	0	2	48‡	0	0	0	0	0	0
Silhouette	0	13	20	16‡	5	0	0	0	0	0
Gap/unif	0	0	0	50‡	1	0	0	0	0	0
Gap/pc	0	0	4	46‡	0	0	0	0	0	0
2 elongated clusters										
CH	0	0‡	0	0	0	0	0	7	16	27
KL	0	50‡	0	0	0	0	0	0	0	0
Hartigan	0	0‡	0	1	0	2	1	5	6	35
Gap/unif	0	0‡	17	16	2	14	1	0	0	0
Gap/pc	0	50‡	0	0	0	0	0	0	0	0

†Numbers are counts out of 50 trials. Some rows do not add up to 50 because the number of clusters chosen was greater than 10.
‡Column corresponding to the correct number of clusters.

Figure 14: Results of the simulation study

8 DISCUSSION

In data that is not clearly separated in groups, people might have different opinions about the number of distinct clusters. When gap statistic is used with a uniform reference distribution in the principal component orientation, it outperforms other proposed methods from the literature. There are many avenues of further research in depth of estimating the number of optimal clusters in a data set. I have created a simple chart to understand the research aspects of gap statistic. Fig.15 shows the mind map of all further research possibilities.

9 ACKNOWLEDGMENTS

I was supported by the DBS Chair at the LMU, Munich with Professor Dr. Thomas Seidl, Collin Leiber and Walid Durani.

10 APPENDICES

[The Proofs of both theorems in the reference distribution are from the paper "**Estimating the number of clusters in a data set via gap statistic**, Robert Tibshirani, Guenther Walther, Trevor Hastie".]

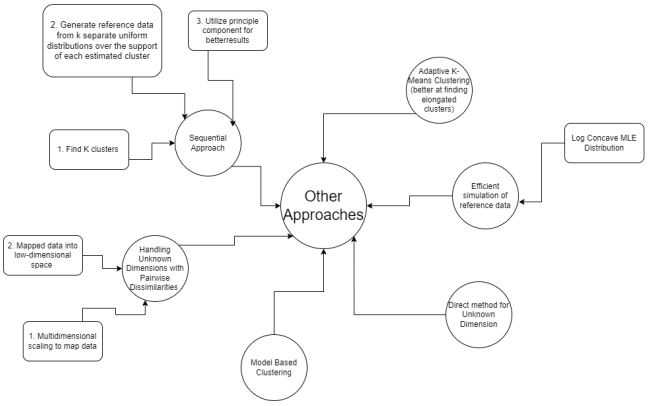

Figure 15: Further Research in gap statistic

REFERENCES

A RESEARCH METHODS

- Robert Tibshirani, Guenther Walther, Trevor Hastie **Estimating the number of clusters in a data set via gap statistic**
- Cuevas, A., Febrero, M. and Fraiman, R. (2000) **Estimating the number of clusters.** *Can. J. Statist.*, ***28****, 367-382*

B ONLINE RESOURCES

- Bradley Bhoehmke **K-Means Cluster Analysis***UC Business Analytics R Programming Guide*
- I.Kabul, P.Hall, J.Silva, W.Sarle **Determining number of clusters in a data set using ABC..** *SAS Institute, MLConf ATL*