Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

12-2022

A recommendation on how to teach K-means in introductory analytics courses

Manoj THULASIDAS
Singapore Management University, manojt@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Higher Education Commons, and the Numerical Analysis and Scientific Computing Commons

Citation

THULASIDAS, Manoj. A recommendation on how to teach K-means in introductory analytics courses. (2022). *2022 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE): Hong Kong, December 4-7: Proceedings*. 46-53.
Available at: https://ink.library.smu.edu.sg/sis_research/7679

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

A Recommendation on How to Teach K-Means in Introductory Analytics Courses

Manoj Thulasidas
School of Computing and Information Systems, *Singapore Management University*, Singapore
manojt@smu.eud.sg, ORCID: 0000-0002-0508-5782

***Abstract*—We teach K-Means clustering in introductory data analytics courses because it is one of the simplest and most widely used unsupervised machine learning algorithms. However, one drawback of this algorithm is that it does not offer a clear method to determine the appropriate number of clusters; it does not have a built-in mechanism for K selection. What is usually taught as the solution for the K Selection problem is the so-called elbow method, where we look at the incremental changes in some quality metric (usually, the sum of squared errors, SSE), trying to find a sudden change. In addition to SSE, we can find many other metrics and methods in the literature. In this paper, we survey several of them, and conclude that the Variance Ratio Criterion (VRC) is an appropriate metric we should consider teaching for K Selection. From a pedagogical perspective, VRC has desirable mathematical properties, which help emphasize the statistical underpinnings of the algorithm, thereby reinforcing the students' understanding through experiential learning. We also list the key concepts targeted by the VRC approach and provide ideas for assignments.**

***Keywords*—K-Means Clustering, Quality Metrics, K Selection, Variance Ratio Criterion**

I. Introduction

K-Means clustering [1] is conceptually simple, and easily explained and understood. For this reason, we teach it to our undergraduate students in their introductory data analytics courses. Practically, however, one of the difficulties that we face in using the algorithm is in selecting the optimal number of clusters to form, which is the so-called "K Selection Problem." We typically teach the elbow-method as the solution, where we chart the variation of the sum of squared errors (SSE) as we vary the number of clusters, and look for a change in its behavior (an "elbow") so as to select the right K. In some instances, we also teach a number of other indexes, which may solve the K selection problem, but with uncertain degrees of success.

In this paper, we survey several of the quality metrics commonly taught and recommended in our classrooms, and systematically study their performance specifically in selecting the right K. We use a wide range of synthetic data as well as some real data sets for this purpose. The aim is to choose the "best" metric based on its performance on real and synthetic data, as well as from a teaching perspective. The right metric should perform well; it should also reinforce statistical concepts and get our students to construct new knowledge, recognizing the connections between the concepts and good practices in data analytics.

II. Related Work

Many of the metrics studied in this paper are also described in an extensive survey [2] of what the the authors called "clustering validity indexes," grouping them in categories. Although it provides valuable information, this survey does not provide a recommendation, especially from a pedagogical perspective.

From a purely pedagogical perspective, it is important to give our students easily interpretable examples to get them to think beyond data [3]. The use of **VRC** as the right tool to use for K Selection, as we recommend, is such an example. This exploration into its theoretical backdrop brings into focus a variety of concepts from statistics [4], perfectly illustrating how they influence and shape our choice. Targeted at computer science and information systems students, we believe this topic will reinforce their understanding and kindle their interest in connecting the concepts in applied statistics to practical applications in real-world data analysis [5].

We will consider and study the performance of seven indexes with a view to finding the best one in a general sense that we can present as the solution to the K selection problem to our students, while also providing a platform to illustrate some of the mathematical ideas behind clustering. In addition to these "classic" quality indexes that we survey in this article, we have several other candidates, some of which are algorithms specifically designed to determine the right K automatically. We have not included them in our study because these approaches are recent research developments, not appropriate in the undergraduate classroom. However, for the sake of completeness, we include their minimal descriptions in this section, considering them related work.

A recent study [6] introduces the Projected Gaussian (PG-Means) method, which performs a K-Means clustering for all Ks in the range of interest and projects both the data and model to a linear subspace. It then looks for a good fit between the model and data using the Kolmogorov-Smirnov (KS) test. PG-Means runs with ten sets of random starting seeds, which our studies indicate may be too small to ensure convergence. The study reports an excellent detection of the number of clusters $K = 20$ on synthetic data with the number of variables $p \in \{2, 4, 8, 16\}$. It also compares PG-Means to X-Means and G-Means [7] and demonstrates its superiority over them.

X-Means [8], originally developed to address the scalability issue of K-Means, also helps determine the right K. An

extension [9] of X-Means is found in the literature, designed to automatically determine K through progressive iterations and merging of clusters based on a BIC stopping rule. G-Means [7] is a method to repeatedly perform K-Means with increasing K until statical tests show that the resulting clusters are Gaussian within a specified confidence level. Other attempts to determine K include a visual assessment of clustering tendency [10].

A recent comparative study [11] argues that relying on any single internal metric or index is unwise, while noting that the WB index [12] (based on sum of squares similar to **VRC**) seems to perform best. Another recent study that defines quality metrics or indexes is a probabilistic approach [13] on external validation of fuzzy clustering, where one data point may belong to multiple clusters. Another one [14] introduces a cluster-level similarity index called the centroid index, focusing on the overall clustering output to quantify the clustering quality. An external quality measure that can apply to many different clustering algorithms, it is not directly comparable to the internal indexes focusing on K-Means. Lastly, in a paper proposal [15], a new separation measure, (termed "dual center") is developed, based on which a validity index is proposed for fuzzy clustering. It is not, however, employed for K selection.

While a comparison to the recent research in the field may be necessary, the focus of this article is on how to introduce the K selection problem and its solution. From this viewpoint, a "Statistical Reasoning Learning Environment," based on the constructivist theory of learning, was suggested [16] about a decade ago to get our students to synthesize new concepts, integrating it with their existing knowledge. The importance of selecting an appropriate metric for K selection, more based on its performance on data rather than simulations, is emphasized in our approach. Such reliance on real-life data is also recommended as having positive effect on student engagement and learning [17] based on a student survey.

III. K-Means Clustering

A. Notations

In order to describe the indexes with consistent notations, we go into some details of the notations in the K-Means algorithm. We start with n observations along p variables, $\vec{\boldsymbol{x}}_i \in \mathbb{R}^p$. The K-Means algorithm will determine K centroids and n cluster assignments such that each observation $\vec{\boldsymbol{x}}_i$ belongs to one and only one cluster.

The cluster assignment is based on the minimum of the Euclidean distances of each observation from all centroids. We will denote the Euclidean distance between two data points $\vec{\boldsymbol{x}}_i$ and $\vec{\boldsymbol{x}}_j$ as $d_{ij}D(\vec{\boldsymbol{x}}_i, \vec{\boldsymbol{x}}_j)$. The cluster assignment of the i^{th} observation is $g_i = \mathbf{argmin}_k d_{ik}$, which returns the k corresponding to the smallest d_{ik}, the distance of the i^{th} observation to the k^{th} cluster centroid ($\vec{\boldsymbol{\mu}}_k$). $d_{ik} = D(\vec{\boldsymbol{x}}_i, \vec{\boldsymbol{\mu}}_k)$.

The centroid $\vec{\boldsymbol{\mu}}_k$ of a cluster is the average of the observations belonging to it. The j^{th} component of the k^{th} centroid is the average of the j^{th} variable of all the data points that belong to the k^{th} cluster: $\mu_{kj} = \mathbf{average}\{x_{ij} : g_i = k\}$, where n_k is the number of observations belonging to the cluster k, and is also referred to as the membership or frequency of the cluster in this paper.

B. Proxies for Clustering Quality

*1) Sum of Squared Errors (**SSE**):* Once converged, K-Means algorithm reports the Sum of Squared Errors, **SSE**. Indeed, most implementations use SSE as the quantity to be minimized to arrive at the best clustering solution. The errors are the distances between each data point and the centroid of the cluster to which it belongs. The distances are squared and summed over all data points to get **SSE**. Some applications report it directly, or after normalizing it using some combination of the number of data points (n) and the number of variables (p). **SSE** is not a dimensionless quantity and is therefore sensitive to the scales or units of the clustering variables.

2) Coefficient of Determination: The Sum of Squared Errors (**SSE**), as defined above, can be thought of as the unexplained variation in the model. Using it in conjunction with the total variation (**SST**) of the clustering variables (obtained from the summary statistics of the data set), we calculate the Coefficient of Determination of the K-Means model as

$$R^2 = 1 - \frac{\mathbf{SSE}}{\mathbf{SST}} \tag{1}$$

Since it is the fraction of the total variation that is explained by the model, R^2 is a ratio and therefore normalizes away some of the inconsistencies related to scale dependence.

3) Within Standard Deviations: Some statistical programs report the so-called "Within" Standard Deviation, which is the standard deviation of the data points belonging to each cluster, computed separately by cluster. As we can see, this entity is closely related to the Sum of Squared Errors (**SSE**) discussed earlier, except that it is reported on a per-cluster basis. Similarly, **SSE** and R^2 may be reported on a per-variable basis as well.

C. Limitations of the Proxies

Both **SSE** and R^2 are susceptible to over-fitting. If we run the K-Means algorithm with as many clusters as data points ($K = n$), for instance, both these proxies will indicate perfect clustering, with $\mathbf{SSE} = 0$, and $R^2 = 1$. In fact, **SSE** usually decreases when we increase K (except for instances where the algorithm finds a local minimum). In order to glean any useful information from it, we use the elbow method, where we try to determine the value of K at which the decrease in **SSE** seems to flatten out. This method is obviously subjective, and not easily automated.

IV. Survey of Indexes

A. Indexes Considered

As mentioned earlier, we have several quality indexes and statistics in the literature, which are frequently used to automatically determine the right number of clusters (K). We list some of the commonly used metrics below along with a short

description of how to use them. Later on, we will benchmark them on our synthetic and real datasets.

VRC The Variance Ratio Criterion (**VRC**) [18] is defined as the ratio of the between and within sums of squares, divided by the appropriate factors to account for the degrees of freedom. The best value for K will correspond to an absolute or local maximum of **VRC** as we vary the number of clusters.

AIC The Akaike Information Criterion [19] is a metric driven by maximum likelihood estimate, and tries to penalize overfitting. The right model (or the right K) is the one that minimizes **AIC**.

BIC The Bayesian Information Criterion [20], another information metric based on maximum likelihood estimate, penalizes complex models even more than **AIC**. Again, the right K would be the one that minimizes **BIC**.

Sil. Wid. The Silhouette visualization [21], originally developed as a graphical display technique, also gives a measure of the clustering quality as the average silhouette width (**Sil. Wid.**), which is a maximum for the best K.

DB The Davies-Bouldin index [22], which computes the dispersion of the output of the clustering algorithm, is a good separation measure. When using this index, one would choose the number of clusters that minimizes the value of **DB** as the right K.

Gap The Gap Statistic [23] uses the output of the clustering algorithm, and compares the within-cluster dispersion against what is expected under an appropriate null reference distribution. The best K recommended by this approach is the smallest number of clusters that shows a decrease in the Gap statistic (corrected internally for the simulation error).

$\boldsymbol{f(K)}$ An evaluation function based on the comparison with a uniform reference [24] (similar to the Gap Statistic), this index is expected to have a better success rate in determining the right number of clusters than **Gap**. All values of K such that $\boldsymbol{f(K)} < 0.85$ are potential candidates as the right K.

B. Quantifying Index Performance

Since we will be comparing multiple indexes, we may get the same right K from several of them. It would then be fair to ask how we quantify the performance of various indexes. For the first five of the seven indexes listed earlier, the selection of K is based on a maximum or a minimum. The last two, **Gap** statistic and the $\boldsymbol{f(K)}$ index, do not determine K by looking for a maximum or minimum in their variation.

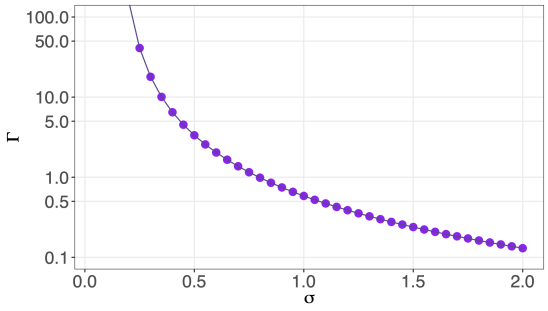


Fig. 1. Computed Γ values (on log scale) for normal distributions of various standard deviations (σ).

For the indexes that determine the ideal K using a maximum or minimum, the significance of K selection may be quantified using the concept of curvature: the higher the curvature, the more prominent the minimum or maximum signifying the right K. Analogously to the three-point computation of the second derivative of a continuous function, for a discrete function $h(K)$ (where K is an integer), we define a new quantity Γ, proportional to the curvature as

$$\Gamma = \left| \frac{h(K+1) - 2h(K) + h(K-1)}{h(K+1) + h(K-1)} \right| \tag{2}$$

The numerator of the expression for Γ is the discrete equivalent of the three-point computation of the second derivative of a continuous function, while the denominator is proportional to the average value of the metric around the K value where the peak is located. Intuitively, it is the ratio of the peak value to the local background average, without counting the peak. Since Γ is dimensionless (with $h(K \pm 1)$ in the denominator), we can think of it as a non-negative quantifier of the quality or the significance of the maximum or minimum. In particular, the value of Γ is unchanged if we scale the metric by a constant.

The lowest value of Γ is 0 when $h(K)$ is flat. In order to get a feel for its meaning, we can compute Γ for normal distributions with several different variances, as shown in Fig 1. For a unit normal distribuiton ($N(\mu, \sigma^2) = N(0, 1)$), $\Gamma = 0.58$. For $\sigma = 0.5$, $\Gamma = 3.34$ and for $\sigma = 0.25$, $\Gamma = 41$. As expected, sharper peaks give larger Γs. We will use Γ when comparing various indices for K selection.

V. Performance on Synthetic Data

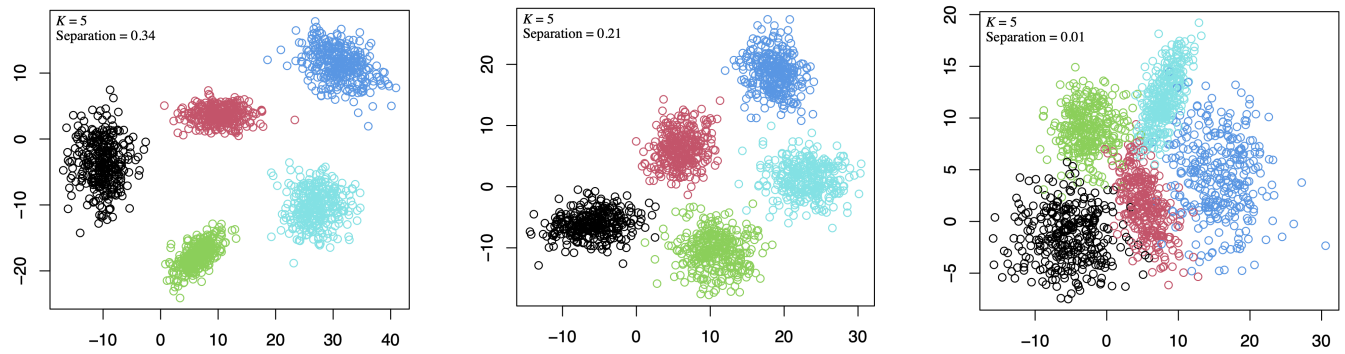
A. Data Generation

Now that we have specified a method to quantify the performance of various quality indexes, we proceed to see how it performs on synthetic data, before moving on to real data. We use the R package `clusterGeneration` [25], which can generate clusters of specified sizes in spaces of prescribed number of variables. In `clusterGeneration`, we can also specify the separation among the clusters, using a separation index [26]. We will use various values for these three and other parameters as described below.

Number of Clusters (G): We generate synthetic data sets with different numbers of clusters: $G \in \{5, 10, 15, 20\}$

Number of Variables (p): For this parameter, we use the values $p \in \{2, 4, 8, 16, 32\}$.

Separation Index (J^*): This parameter controls how well separated the clusters are, and has range $-1 < J^* < 1$. We use three values for the separation index:
$J^* = 0.34$, which gives cleanly separated clusters,
$J^* = 0.21$, which gives relatively clean clusters, and
$J^* = 0.01$, which gives realistic clusters.
The effect of changing J^* is shown in Fig 2.



(a) Well-separated Clusters ($J^* = 0.34$) (b) Medium separation ($J^* = 0.21$) (c) Realistic Clusters ($J^* = 0.01$)

Fig. 2. Sample of clusters ($K = 5, p = 2$) generated in the synthetic data for various values of the separation index J^*.

With these settings, we have 60 sets of synthetic data, 20 of which have a separation index $J^* = 0.34$. Since we are studying the metric for a data set perfectly suited for K-Means clustering, we focus on these 20 data sets for detailed analysis, while providing summary comparisons on all 60. Keeping the same rationale of balanced and distinct clusters in mind, we use the following values for the other parameters in the generation of the synthetic data.

- Number of noisy variables $= 0$
- Number of outliers $= 0$
- Equal cluster membership (of $10p$) for all clusters
- Cluster uniformity (= Range for variances of the covariance matrix) $= [1, 10]$, which generates a reasonable variability.

B. Analysis of Synthetic Data

With the synthetic data generated, we can study the performance of the seven indexes (**VRC**, **AIC**, **BIC**, **Sil. Wid.**, **DB**, **Gap** and $\boldsymbol{f(K)}$) discussed earlier.

For each run of the K-Means clustering algorithm, we use 100 random sets of initial seeds from which the best run (based on the sum of squared errors) is chosen. It is important to have large number of starting seeds because of the sensitivity of K-Means to initial conditions, especially when we have large number of clusters and relatively small number of variables [27]. For smaller number of starting seeds, we do see a large fraction of K-Means attempts failing to converge. We also set a generous limit on the maximum number of iterations of 1000. Still concerned about the robustness of the analysis, we repeat the whole analysis multiple times and ensure that the results reported are stable.

C. Results and Discussion

After we run the analysis on all our 60 synthetic data sets (20 for each J^* value), we report the average accuracy for the indexes in Table I. Also reported are the average Γ values when the right K is detected. Since we are looking for a metric that will work best for data sets that are particularly suited for K-Means clustering, the column to consider in Table I is for well-separated clusters ($J^* = 0.34$). The Variance Ratio Criterion (**VRC**) and the Davies-Bouldin index (**DB**) detect the right K. The significance of the peak **VRC** for and of the minimum for **DB** is at similar levels ($\Gamma \approx 0.1 - 0.2$).

It is noteworthy that **VRC**, **DB** and **Gap** perform well when the clusters are generated with overlaps, with $J^* = 0.01$, when the clusters are expected to be realistic. Although these indexes have high accuracy in the synthetic data, their performance in the real data is not impressive.

Another important point to note is that both **AIC** and **BIC** perform very poorly in both the synthetic data and the real data, as described in the following section. One of the recommendations of this paper is to avoid teaching these information criteria as a K selection method.

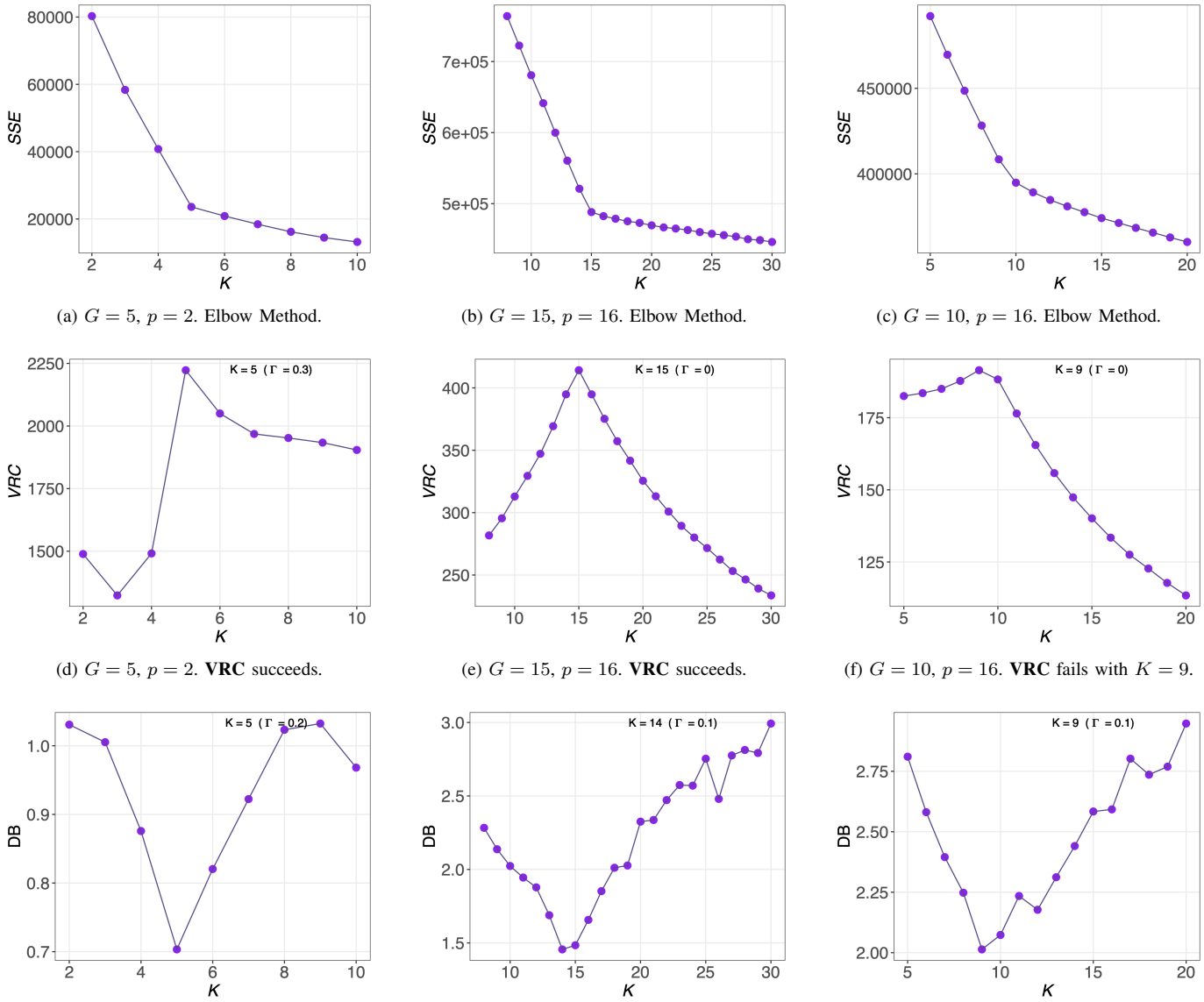
VI. Performance on Real Data

We also perform our comparison in four different real data sets. We detect the optimal number of clusters automatically using the indexes, and compare it to what is known about the data sets independently. We first outline the variable selection applied to all four data sets. We then proceed to describe the data sets and present the results of our studies. In these studies, we assume that the classes in the data sets form spherical clusters, easily separated by the K-Means algorithm, and, as a

TABLE I
Accuracy and Γ of various metrics

Metric	Well-separated ($J^* = 0.34$)	Medium ($J^* = 0.21$)	Realistic ($J^* = 0.01$)
VRC	100.0% (0.2)	100.0% (0.2)	60.0% (0.1)
AIC	0.0% (−)	0.0% (−)	0.0% (−)
BIC	0.0% (−)	0.0% (−)	0.0% (−)
Sil. Wid.	95.0% (0.1)	75.0% (0.1)	65.0% (0.1)
DB	100.0% (0.2)	85.0% (0.2)	85.0% (0.1)
Gap	55.0%	70.0%	70.0%
$\boldsymbol{f(K)}$	20.0%	25.0%	30.0%

Accuracy of K selection: the fraction of the times when the reconstructed number of clusters is the same as the generated number ($K = G$). The numbers between parentheses are the mean Γ (averaged when $K = G$).



(a) $G = 5$, $p = 2$. Elbow Method.
(b) $G = 15$, $p = 16$. Elbow Method.
(c) $G = 10$, $p = 16$. Elbow Method.
(d) $G = 5$, $p = 2$. **VRC** succeeds.
(e) $G = 15$, $p = 16$. **VRC** succeeds.
(f) $G = 10$, $p = 16$. **VRC** fails with $K = 9$.
(g) $G = 5$, $p = 2$. **DB** succeeds.
(h) $G = 15$, $p = 16$. **DB** fails with $K = 14$.
(i) $G = 10$, $p = 16$. **DB** fails with $K = 9$.

Fig. 3. Examples of the Elbow method using **SSE** in the top row, the Variance Ratio Criterion (**VRC**) in the second row and the Davies-Bouldin index **DB** in the bottom row for different number of clusters generated G with the specified number of variables (p). The first and last columns are with the separation value $J^* = 0.01$ while the middle column is with $J^* = 0.21$.

consequence, that the ideal number of clusters is the number of classes.

A. Variable Selection

Since we are testing the indexes on labeled data sets, we can directly compute the purity of the clusters by counting the number of correctly assigned observations. We assume that the ideal number of clusters (ideal K) is the number of distinct values of the label. After a clustering run with the ideal K, we select the variables that give the highest purity. For the set of clusters $\Omega = \{\omega_1, \omega_2, \cdots, \omega_k\}$ and the set of classes $\mathbb{C} = \{c_1, c_2, \cdots, c_k\}$, purity is defined as

$$\text{Purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{i=1}^{K} \max_j (\omega_i \cap c_j) \tag{3}$$

For each cluster ω_i, we find the most frequent class label c_j, and compute the purity as the fraction of the sum of the number of most frequent labels. Purity can be easily calculated from the confusion matrix, as illustrated in Table II for the Iris data set. We can see that the species *Setosa* has been clustered perfectly into Cluster 2, while *Versicolor* is mostly in 1 and *Virgnica* mostly in 3. Purity is the sum of the largest values in each row ($50 + 48 + 36 = 134$) divided by the total number of observations ($n = 150$) $= 89.3\%$.

We then repeat the computation of purity for all possible combinations of variables. In the Iris data set, we have 11 such combinations. If there are multiple rows with the same highest purity, we select the combination with the highest χ^2_R, which gives us the best variables to use in clustering the data set. As we can see, the best variable combination to use for the Iris data set would be *Petal Length* and *Petal Width*. Note that we will use the same "best" variables for all other indexes.

B. Data Sets

1) Iris Data Set: The classic Iris data set [28] contains 150 flower measurements along four variables (*Sepal Length*, *Sepal Width*, *Petal Length* and *Petal Width*) from three different iris species (*Setosa*, *Versicolor* and *Virgnica*). Each species has 50 data points in the data set. Since there are three species, we know, beforehand, that the ideal number of clusters should be three. As described earlier, we select the variables *Petal Length* and *Petal Width* as the variables to use when looking for the best K.

2) Young Adults Data Set: We collected anonymous data from our students. The data set has 127 observations of four numeric variables (*Height*, *Weight*, *Age* and *HairLength*) and a label (M or F for male or female). Note that in Singapore, male university students are expected to be about 2 to 3 years older than their female classmates because of their military service obligation. Therefore, we may expect the *Age* variable to have some differentiating power while clustering the data. Following the same procedure as in the iris data set, we select *Weight* and *HairLength* as the best variables to use, for the best possible purity of 98.4%. From our purity studies, the *Age* variable does not seem to contribute in segregating the classes.

3) The Wine Data Set: The publicly available Wine data set [29], from the UCI Machine Learning Repository [30], has 12 attributes, making the combinatorial problem of selecting the best variables for K-Means clustering challenging with over 8000 possible combinations. From among the multiple variable combinations that give the highest purity of 90.5%, we select the combination of *Alcohol*, *Ash*, *Flavanoids* and *OD280_OD315* based on the highest purity. The Wine data set also has three classes.

4) The Seeds Data Set: The publicly available Seeds data set [31] (again from the UCI Machine Learning Repository) contains three classes of wheat seeds with 70 observations each. It has seven attributes, giving us 120 different combinations of variables to choose from. The highest purity that we can get as we use different combinations of variables is 90.0%. From these combinations, we select *Area*, *Perimeter*, *Compactness* and *Asymmetry* based on the highest purity.

TABLE II
COMPUTING PURITY FROM THE CONFUSION MATRIX

Species	Cluster 1	Cluster 2	Cluster 3
Setosa	0	50	0
Versicolor	48	0	2
Virgnica	14	0	36

Number of observations with different labels and in various clusters for the Iris data set. We can see that the species *Setosa* has been clustered into Cluster 2, *Versicolor* in 1 and *Virgnica* in 3.

TABLE III
COMPARISON OF INDEXES

	Data Set			
Index	Iris	YA	Wine	Seeds
G	**3**	**2**	**3**	**3**
VRC	10 (0.03)	**2 (0.08)**	**3 (0.25)**	**3 (0.11)**
AIC	5 (0.04)	12 (−)	6 (0.01)	12 (−)
BIC	4 (0.05)	12 (−)	**3 (0.17)**	9 (0.01)
Sil. Wid.	2 (0.16)	**2 (0.12)**	**3 (0.16)**	2 (0.09)
DB	2 (0.41)	**2 (0.20)**	**3 (0.22)**	2 (0.10)
Gap	5	3	**3**	**3**
$\boldsymbol{f(K)}$	2	**2**	2	2

Comparison of the performance of various commonly used quality indexes. The top row is G, the number of classes in our data sets. When an index predicts the right K, it is highlighted in **bold**. (Γ, defined in Eq. (2), the significance of the peak, and is reported between parentheses. Γ cannot be calculated when $K = 12$ because it is the end of the K range.)

We can see from Table III that of the seven other indexes considered, the **VRC** index seems to perform best with three right predictions.

VII. Recommendations, Concepts and Assignments

A. What Index to Teach

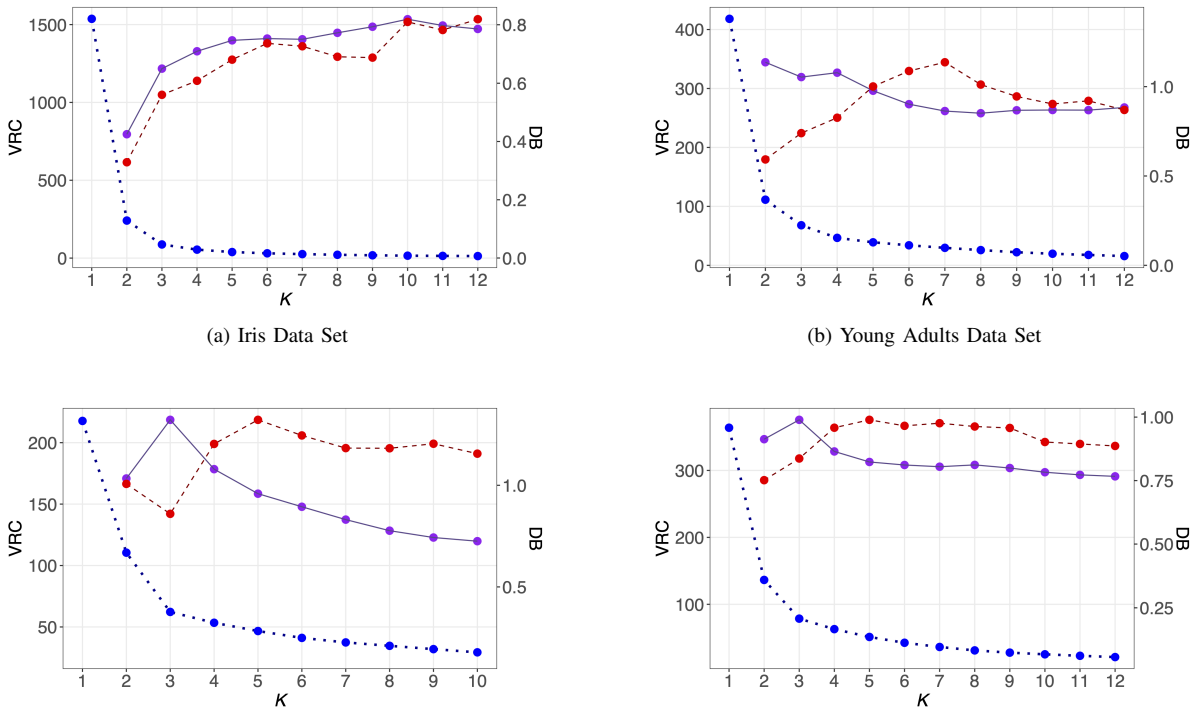
From our studies, it seems clear that **VRC** is the right index to use to solve the K selection problem. The popular "elbow" method, which looks for a kink in the variation of the sum of squared errors, and is subjective and impossible to automate.

B. How to Teach *VRC*

The Variance Ratio Criterion (**VRC**) is the ratio of the inter-cluster variance to intra-cluster variance (hence the name "Variance Ratio"). The variances are computed from the sum of squared errors, with the appropriate statistical normalization factors. **VRC**, therefore, has a remarkably intuitive description. In the original paper [18], they used the variables between-sum-of-squared-errors (**BGSS**) and within-sum-of-squared-errors (**WGSS**). Using our notations, we can rewrite **BGSS** and **WGSS** as:

$$\begin{aligned}
\mathbf{SST} &= \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\mu_j)^2 \\
\mathbf{WGSS} = \mathbf{SSE} &= \sum_{i:g_i=k}^{n}\sum_{j=1}^{p}\left(x_{ij}-\mu_{k_j}\right)^2 \\
\mathbf{BGSS} &= \mathbf{SST}-\mathbf{SSE} \\
\mathbf{VRC} &= \frac{\mathbf{BGSS}}{K-1}\Big/\frac{\mathbf{WGSS}}{n-K}
\end{aligned} \tag{4}$$

Note that we used **SST** and **SSE** in the definition of R^2 in Eq. (1). As we can see, **VRC** is closely related to R^2 and the



(a) Iris Data Set

(b) Young Adults Data Set

(c) Wine Data Set

(d) Seeds Data Set

Fig. 4. Using **VRC** (solid line with purple points, with scale on the left) and **DB** (dashed line with red points, with scale on the right) to select the optimal number of clusters in various data set. Also shown is the shape elbow-plot of **SSE** (blue dotted line, without a scale). As summarized in Table III, VRC succeeds in getting the right K in three out of the four data sets.

F-statistic, which makes it easier to teach it in introductory data science courses to undergraduate students as a way to reinforce their statistical thinking. For instance, we can rewrite **VRC** in terms of R^2 using Eq. (1) as:

$$\textbf{VRC} = \frac{R^2}{1-R^2}\,\frac{n-K}{K-1} \tag{5}$$

VRC is easily computed using the quantities reported by most K-Means clustering implementations. In `R`, for instance, the `kmeans` function (from the basic `stats` package) reports **BGSS** as `betweenss` and **WGSS** as `tot.withinss`.

SAS reports R^2 and also the square roots of **BGSS** and **WGSS** (after scaling them by the appropriate statistical factors) as `Cluster Standard Deviations` and `Statistics of Variables`.

C. Key Concepts

In addition to the formal descriptions of the K-Means and k-NN, the inclusion of **VRC** in the discourse brings a host of concepts and practices that can be taught in classrooms in an experiential learning framework.

Statistical Results: The statistical entities reported by K-Means applications, namely R^2, **SST** and **SSE**, are used in the computation of **VRC** (Equations (4) and (5)), which gives us a good opportunity to highlight their relevance.

Degrees of Freedom: The **VRC** computation in Equations (4) and (5) also illustrates the difficult concept of degrees of freedom and the need to carefully normalize statistical entities for comparison purposes.

Limitations of Algorithms: The implicit assumptions in K-Means (Fig 2), not often highlighted in the classroom, can be used to demonstrate the need to diligently explore the usability conditions of algorithms in general.

Feature Selection: The variable selection using purity described in the article can be used as a tool to teach feature selection techniques in labeled data.

Curvature for Index Selection: The key concept of calculus (often not relevant in data analytics) can be used to quantify the significance of the peak in a quantity of interest, as in Eq. (2).

Monte Carlo Simulation: In Section §V-A, we provide step-by-step instructions on how to generate and use synthetic data from a multivariate normal distribution with a specified covariance matrix and means.

Automated Insight Generation: An advantage of **VRC** as the solution to the K Selection problem is that its use can be automated (because it is a straightforward search for a maximum), thereby bringing the K-Means algorithm more in line with other tools in data analytics, capable of automatic generation of insights.

D. Assignment Ideas

Here are some ideas for formative assessments when the NC algorithm is used as a topic in a course. The solutions

and discussions of these ideas are available from the authors upon request.

- Do a Monte Carlo simulation to generate synthetic data and reproduce the plots given in Fig 3.
- Establish the validity of the Γ method, described in Eq. (2), for index comparison using simulation.
- Compare the performance of various indexes in detecting the right number of clusters to form.
- Explore ways in which the elbow method can be automated.

VIII. Conclusion

We studied various indexes in their accuracy in detecting the right number of clusters to form in K-Means clustering, with a view to recommending the right method to teach in introductory data analytics courses. In addition to the accuracy, we also considered the relevance of the key learning points and other pedagogical advantages such as the connection to statistical concepts.

From our comparative studies on synthetic data, we see that the Variance Ratio Criterion (**VRC**) and the Davies-Bouldin index (**DB**) work remarkably well in the synthetic data sets. In real data, however, **VRC** outperforms **DB**. Moreover, the significance of the peak indicating the right K seems larger for **VRC** compared to **DB**. All other indexes performed poorly both on both synthetic as well as real data. In particular, we conclude that both the Akaike and the Bayesian Information Criteria (**AIC** and **BIC**) are ineffectual in selecting the right K in K-Means clustering. The Gap Statistic performs slightly better than the information criteria, but it is prohibitively expensive, computationally.

From the perspectives of automation, simplicity and comprehensibility, the Variance Ratio Criterion is perhaps the best index to teach our undergraduate students embarking on their data science curriculum. With its connection to familiar concepts in statistics (such as sum of squared errors, coefficient of determination etc.), **VRC** can be taught as an extension of existing knowledge and beliefs that the students already have, thus facilitating the creation, integration and assimilation of new ideas.

References

[1] J. A. Hartigan, *Clustering algorithms*, ser. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1975.

[2] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243 – 256, 2013.

[3] K. Lübke, M. Gehrke, J. Horst, and G. Szepannek, "Why we should teach causal inference: Examples in linear regression with simulated data," *Journal of Statistics Education*, vol. 28, pp. 1–17, 04 2020.

[4] B. Chance, "Components of statistical thinking and implications for instruction and assessment," *Journal of Statistics Education*, vol. 10, 11 2002.

[5] J. Singer and J. Willett, "Improving the teaching of applied statistics: Putting the data back into data analysis," *American Statistician*, vol. 44, pp. 223–230, 08 1990.

[6] Y. Feng and G. Hamerly, "Pg-means: learning the number of clusters in data," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006.

[7] G. Hamerly and C. Elkan, "Learning the k in k-means," *Advances in Neural Information Processing Systems*, vol. 17, 03 2004.

[8] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *ICML*, 2000.

[9] T. Ishioka, "An expansion of x-means for automatically determining the optimal number of clusters," in *Computational Intelligence*, 2005.

[10] M. Pakhira, "Finding number of clusters before finding clusters," *Procedia Technology*, vol. 4, pp. 27–37, 12 2012.

[11] J. Hämäläinen, S. Jauhiainen, and T. Kärkkäinen, "Comparison of internal clustering validation indices for prototype-based clustering," *Algorithms*, vol. 10, p. 105, 09 2017.

[12] Q. Zhao and P. Fränti, "Wb-index: A sum-of-squares based index for cluster validity," *Data & Knowledge Engineering*, vol. 92, pp. 77 – 89, 2014.

[13] D. Campo, G. Stegmayer, and D. Milone, "A new index for clustering validation with overlapped clusters," *Expert Systems with Applications*, vol. 64, pp. 549 – 556, 2016.

[14] P. Fränti, M. Rezaei, and Q. Zhao, "Centroid index: Cluster level similarity measure," *Pattern Recognition*, vol. 47, no. 9, pp. 3034 – 3045, 2014.

[15] S. Yue, J. Wang, J.-S. Wang, and X. Bao, "A new validity index for evaluating the clustering results by partitional clustering algorithms," *Soft Computing*, vol. 20, pp. 1127–1138, 2016.

[16] J. Garfield and D. Ben-Zvi, "Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment," *Teaching Statistics*, vol. 31, pp. 72 – 77, 08 2009.

[17] D. Neumann, M. Hood, and M. Neumann, "Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course," *Statistics Education Research Journal*, vol. 12, pp. 59–70, 11 2013.

[18] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-Simulation and Computation*, vol. 3, pp. 1–27, 1974.

[19] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[20] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[21] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.

[22] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, pp. 224–227, 1979.

[23] R. Tibshirani, W. Guenther, and H. Trevor, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, pp. 411–423, 2002.

[24] D. Pham, S. Dimov, and C. Nguyen, "Selection of K in K-means clustering," *Proceedings of The Institution of Mechanical Engineers Part C-journal of Mechanical Engineering Science - PROC INST MECH ENG C-J MECH E*, vol. 219, pp. 103–119, 2005.

[25] W. Qiu and H. Joe, "Generation of random clusters with specified degree of separation," *Journal of Classification*, vol. 23, no. 2, pp. 315–334, Sep. 2006.

[26] ——, "Separation index and partial membership for clustering," *Computational Statistics & Data Analysis*, vol. 50, pp. 585–603, 02 2006.

[27] S. Sieranoja, "How much k-means can be improved by using better initialization and repeats?" *Pattern Recognition*, vol. 93, 04 2019.

[28] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[29] S. Aeberhard, D. Coomans, and O. de Vel, "Comparison of classifiers in high dimensional settings," Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland, Tech. Rep. 92-02, 1992.

[30] D. Dheeru and E. K. Taniskidou, "UCI machine learning repository," 2017.

[31] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak, "Complete gradient clustering algorithm for features analysis of x-ray images," *Information Technologies in Biomedicine*, vol. 69, pp. 15–24, 2010.