

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

11-2022

A quality metric for K-Means clustering based on centroid locations

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 [Computer Engineering Commons](#), [Numerical Analysis and Scientific Computing Commons](#), and the [Theory and Algorithms Commons](#)

Citation


THULASIDAS, Manoj. A quality metric for K-Means clustering based on centroid locations. (2022). *Advanced Data Mining and Applications: 18th International Conference, ADMA 2022, Brisbane, Australia, November 28-30: Proceedings*. 13726, 208-222.

Available at: https://ink.library.smu.edu.sg/sis_research/7744

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 Quality Metric for K-Means Clustering Based on Centroid Locations

Manoj Thulasidas^(✉) 

School of Computing and Information Systems, Singapore Management University,
80 Stamford Road, Singapore 178902, Singapore
manojt@smu.edu.sg

Abstract. K-Means clustering algorithm does not offer a clear methodology to determine the appropriate number of clusters; it does not have a built-in mechanism for K selection. In this paper, we present a new metric for clustering quality and describe its use for K selection. The proposed metric, based on the locations of the centroids, as well as the desired properties of the clusters, is developed in two stages. In the initial stage, we take into account the full covariance matrix of the clustering variables, thereby making it mathematically similar to a reduced χ^2 . We then extend it to account for how well the clustering results comply with the underlying assumptions of the K-Means algorithm (namely, balanced clusters in terms of variance and membership), and define our final metric (\mathcal{M}_C). We demonstrate, using synthetic and real data sets, how well our metric performs in determining the right number of clusters to form. We also present detailed comparisons with existing quality indexes for automatic determination of the number of clusters.

Keywords: K-Means clustering · Quality metrics · K selection problem · Number of clusters

1 Introduction

K-Means clustering [15] is conceptually simple and easily explained and understood. Practically, however, one of the difficulties that we face in using the algorithm is that we cannot clearly and objectively articulate why one clustering output is better than another one for a given data set. We lack a quality measure. Because of the lack of a quality measure, we face difficulties when it comes to selecting the optimal number of clusters to form.

In this paper, we propose a new quality metric that can be easily computed during (or after) K-Means clustering and argue from basic principles that it accurately captures the validity of the clustering run. We will study its performance in determining the optimal number of clusters to form on a wide range of synthetic data as well as some real data sets. We will demonstrate that it compares favourably against the current metrics, several of which are reviewed in [4].

2 Related Work

We have several quality indexes and statistics in the literature, which are frequently used to automatically determine the right number of clusters (K). The ones we will consider in the article for comparison are:

- Variance Ratio Criterion [5]: **VRC**
- Akaike Information Criterion [3]: **AIC**
- Bayesian Information Criterion [24]: **BIC**
- Silhouette Width [23]: **Sil. Wid.**
- Gap Statistic [26]: **Gap**
- Evaluation Function [20]: $f(K)$

In addition to these “classic” quality indexes, we have several other candidates, some of which are algorithms specifically designed to determine the right K automatically. A recent study [10] introduces the Projected Gaussian (PG-Means) method, which performs a K-Means clustering for all K s 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.

X-Means [19], originally developed to address the scalability issue of K-Means, also helps determine the right K . An extension [16] 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. This method, however, does not give an index, which is needed for other purposes such as feature selection.

G-Means [14] is a method to repeatedly perform K-Means with increasing K until statistical tests show that the resulting clusters are Gaussian within a specified confidence level. This method again does not provide a quality metric. Other attempts to determine K include a visual assessment of clustering tendency [18], again with no overall quality metric.

A recent comparative study [13] argues that relying on any single internal metric or index is unwise, while noting that the WB index [28] (based on sum of squares similar to **VRC**) seems to perform best. Our index, also loosely based on sum of squares, seems to work well both in synthetic and real data sets.

One of the more recent studies that define quality metrics or indexes is a probabilistic approach [6] on external validation of fuzzy clustering, where one data point may belong to multiple clusters. Our approach also uses within-standard deviations, and applies only to K-Means clustering, which is distinctly non-fuzzy. Another approach [12] 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 our internal metric focusing on K-Means. Lastly, in a paper proposal [27], 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.

3 New Quality Metrics

To develop the metric proposed in this article, we will start from the standard (z) scores of the centroid locations and combine them into a metric. We will then generalize it using the full covariance matrix of the clustering variables (grouped by cluster) to define a reduced χ^2 metric. At the second stage, we will extend the χ^2 metric to incorporate extra information about how well the clusters conform to the implicit assumptions in the K-Means algorithm and come up with the proposed metric, \mathcal{M}_C .

Given the centroids ($\vec{\mu}_k$) and the population mean ($\vec{\mu}$), we can compute the significance of the difference between them for each variable as,

$$z_{k_j} = \frac{\delta_{k_j}}{\sigma_{k_j}^{(c)}} = \frac{\delta_{k_j}}{\frac{\sigma_{k_j}}{\sqrt{n_k}}} = \frac{\sqrt{n_k}(\mu_j - \mu_{k_j})}{\sigma_{k_j}} \quad (1)$$

where δ stands for the difference and $\sigma^{(c)}$ for the within-cluster standard deviation. Since the k^{th} cluster has n_k members, the standard error is $\sigma^{(c)}$ divided by $\sqrt{n_k}$. In order to interpret the squared sum as a weighted average, we divide it by the number of observations n , so that each term in the sum has a weight of n_k/n , the fraction of the observations belonging to the cluster, and call it our quality **Score**.

$$\begin{aligned} \text{Score} &= \frac{1}{K(p-1)} \sum_{k=1}^K \frac{|\vec{z}_k|^2}{n} \\ &= \frac{1}{nK(p-1)} \sum_{k=1}^K n_k \sum_{j=1}^p \left(\frac{\mu_j - \mu_{k_j}}{\sigma_{k_j}} \right)^2 \end{aligned} \quad (2)$$

3.1 Reduced χ_R^2 Metric

The generalized version of the distance to be used in the presence of correlations is the Mahalanobis Distance [17], $D_M(\vec{\mu}_k, \vec{\mu})$ corresponding to the K cluster centroids. The square of each one (denoted by $D_M^2(\vec{\mu}_k, \vec{\mu})$) is a random variable which follows a χ^2 distribution with a parameter (or degrees of freedom, **DoF**) $p-1$, where p is the number of clustering variables. We can combine these Mahalanobis distances in quadrature using the same weightage as in the definition of **Score**.

$$\begin{aligned} \chi_R^2 &= \frac{1}{K(p-1)} \sum_{k=1}^K \frac{D_M^2(\vec{\mu}_k, \vec{\mu})}{n} \\ &= \frac{1}{nK(p-1)} \sum_{k=1}^K n_k (\vec{\mu}_k - \vec{\mu}) \Sigma_k^{-1} (\vec{\mu}_k - \vec{\mu})^\top \end{aligned} \quad (3)$$

The sum of the squares of the K Mahalanobis distances, being the sum of K random variables, each with a χ^2 (of **DoF** = $p-1$) distribution, is another χ^2 random variable of **DoF** = $K(p-1)$.

Since $K(p - 1)$ is actually the number of degrees of freedom, χ_R^2 can be thought of as the reduced χ^2 per cluster, but with an extra (constant) scaling factor of n . This scaling, being constant, does not impact the usage of χ_R^2 in determining the right K . We call this reduced and *scaled* χ_R^2 our “Reduced χ_R^2 Metric.”

3.2 Implicit Assumptions in K-Means Algorithm

The K-Means algorithm works best when the data set has spherical clusters of roughly equal sizes. The clusters are expected to be similar in terms of membership, density and variance. If this assumption is violated, the K-Means algorithm is likely to give unreliable results. Furthermore, if one cluster has significantly smaller variance or number of members, it tends to “scavenge” observations belonging to other clusters. This is because the cluster boundaries are perpendicular bisectors and the statistical fluctuations in the observations always favor the tighter or smaller cluster. The soft requirement of balanced clusters in terms of membership and variance forms an implicit assumption in the algorithm.

3.3 Covariant Metric (\mathcal{M}_C)

Since the metric is a reduced χ^2 , it may be possible extend it to include components that quantify these assumptions in the K-Means algorithm. We will show how the cluster membership (or frequency) and the cluster standard deviation are compared against their expected or ideal values, and a standard score for each is generated, to be combined with χ_R^2 . We will call the extended metric the Covariant Metric (\mathcal{M}_C) because it is built on the covariance matrix of the data. We emphasize that it is weighted by n_k/n and therefore does not numerically equal standard score or the reduced χ^2 , and it incorporates the components described below in a heuristic way.

Cluster Frequency. Since we have n observations and K clusters, the “ideal” frequency for each cluster is $\hat{n}_k = n/K$. Assuming Poisson statistics, we can argue that the expected error on each frequency is $\sqrt{n/K}$. Since we have K measurements of the frequencies, we gain another factor of \sqrt{K} in its standard error, giving us $\sigma_{n_k} = \frac{\sqrt{n}}{K}$. and combine the individual z-scores in quadrature to come up with a measure of how far away our clustering result is from the ideal, in terms of the membership frequency.

$$M_{n_k} = \sum_{k=1}^K \left(\frac{n_k - \hat{n}_k}{\sigma_{n_k}} \right)^2 \quad (4)$$

M_{n_k} is a standardized measure of how different the clusters are in terms of their frequency. Ideally, we would like to have M_{n_k} as close to zero as possible.

Cluster Variance. Once the clustering is done, we have the sum of squared errors **SSE**. If the clusters have the same variance, then **SSE** should be shared among them in proportion to the frequency.

$$\mathbf{SSE}_k = \frac{n_k - 1}{n - K} \mathbf{SSE} \quad (5)$$

\mathbf{SSE}_k is the sum of the squared errors of the observations to their respective centroids. The expected “ideal” variance, therefore, is this sum divided by $n_k - 1$.

$$\hat{S}_k^2 = \frac{\mathbf{SSE}_k}{n_k - 1} = \frac{\mathbf{SSE}}{n - K} \quad (6)$$

The actual variances of the clusters are estimated during the clustering process, and is reported in terms of within standard deviations, but aggregated over all variables. Ignoring the cases where $n_k = 1$,

$$S_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n;g_i=k} \sum_{j=1}^p (x_{ij} - \mu_{k_j})^2 \quad (7)$$

The standard error in the variance is obtained by recognizing that the sample variance (when multiplied by $(n_k - 1)/\sigma_{S_k^2}^2$) is a χ^2 distribution of $n_k - 1$ degrees of freedom, which itself has a variance of $2(n_k - 1)$. Therefore, the standard error of the variance is [2]

$$\sigma_{S_k^2} = S_k^2 \sqrt{\frac{2}{n_k - 1}} \quad (8)$$

Again, we have an “ideal” variance and a measured one, and we can compute the significance of the difference between them (using the standard errors) and combine their significances to come up with a measure of how the cluster variances compare to the ideal equal variance.

$$M_{S_k^2} = \sum_{k=1}^K \left(\frac{S_k^2 - \hat{S}_k^2}{\sigma_{S_k^2}} \right)^2 \quad (9)$$

In an ideal clustering solution, we will expect to have very small $M_{S_k^2}$.

Extending the χ_R^2 Metric. Now that we have the two new components encapsulating the uniformity among the clusters in terms of frequency and variance, we can extend our χ_R^2 with them to obtain the Covariant Metric (\mathcal{M}_C) as follows.

$$\mathcal{M}_C = \frac{\chi_R^2}{M_{n_k} + M_{S_k^2}} \quad (10)$$

where M_{n_k} and $M_{S_k^2}$ are defined above in Eqs. (4) and (9) above. We divide by the sum of these two measures corresponding to the frequencies and variances of the clusters because the overall quality of the K-Means clustering is inversely proportional to them. In other words, if we have two clustering solutions with

identical χ_R^2 , but different values for M_{n_k} and $M_{S_k^2}$, we have to choose the one with the lower M_{n_k} and $M_{S_k^2}$. Note, however, that a more general way to combine them would be as a linear combination, $w_1 M_{n_k} + w_2 M_{S_k^2}$, where w_1 and w_2 are relative weights whose values are not known a priori.

3.4 Quantifying Index Performance

Since we will be comparing multiple indexes with our metric, 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 out of the seven indexes listed earlier (namely **VRC**, **AIC**, **BIC**, **Sil. Wid.** and **DB**), the selection of K is based on a maximum or a minimum. The **Gap** statistic and the $f(\mathbf{K})$ index do not determine K by looking for a maximum or minimum in their variation.

For the **Gap** statistic, the best K recommended by this approach is the smallest number of clusters that shows a decrease, while all values of K such that $f(\mathbf{K}) < 0.85$ are potential candidates as the right K .

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 . For a continuous function of a single variable, the curvature is proportional to the second derivative. For a discrete function $h(K)$ (where K is an integer), we define a new quantity Γ , similar to the three-point computation of the second derivative for a continuous functions.

$$\Gamma = \left| \frac{h(K+1) - 2h(K) + h(K-1)}{h(K+1) + h(K-1)} \right| \quad (11)$$

The index with the largest Γ value has the most clearly defined peak, signifying the right K .

4 Experiments on Synthetic Data

4.1 Data Generation

We use the R package `clusterGeneration` [21], 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 [22]. 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): We use the values $p \in \{2, 4, 8, 16, 32\}$ for this parameter

Separation Index (J^*): This parameter controls how well separated the clusters are, and we use the value $J^* = 0.34$ (for cleanly separated clusters), since we are defining and studying the metric for a data set well suited for K-Means clustering.

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, 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.

4.2 Analysis of Synthetic Data

With the synthetic data, we first compute our metrics \mathcal{M}_C , χ_R^2 , and seven indexes (**VRC**, **AIC**, **BIC**, **Sil. Wid.**, **DB**, **Gap** and $f(\mathbf{K})$) 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 [25]. 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 and repeat the whole analysis multiple times and ensure that the results reported are stable.

4.3 Results and Discussion

First, we focus on the fraction of the times we can detect the right number of clusters using the metrics \mathcal{M}_C , χ_R^2 , **VRC**, **AIC**, **BIC**, **Sil. Wid.**, **DB**, **Gap** and $f(\mathbf{K})$. We define this fraction as the accuracy of the metric and scan for the right K in $\frac{G}{2} < K < 2G$. We run the analysis on all our 60 synthetic data sets (20 for each J^* value), and report the average accuracy for our metric and a variety of indexes in Table 1. Also reported are the average Γ values when the right K is detected. Since we are developing a metric that will work best for data sets that are particularly suited for K-Means clustering, the column to consider in Table 1 is for well-separated clusters ($J^* = 0.34$). We can see that \mathcal{M}_C performs very well with extremely well-defined peaks ($\Gamma \approx 28$). Although the Variance Ratio Criterion (**VRC**) and the Davies-Bouldin index (**DB**) also detect the right K , the significance of the peak for **VRC** or the minimum for **DB** is at much smaller levels ($\Gamma \approx 0.1 - 0.2$).

VRC, **DB** and **Gap** perform better than \mathcal{M}_C when the clusters are generated with more overlaps, by reducing the value of J^* to 0.01, when the clusters are expected to be more realistic. However, their performance in the real data is poorer than our metric. Also of note is that both **AIC** and **BIC** perform very poorly in the synthetic data, as well as in the real data. (The comparisons of the performances of the metrics in real data is summarized in Table 2 in a subsequent section.)

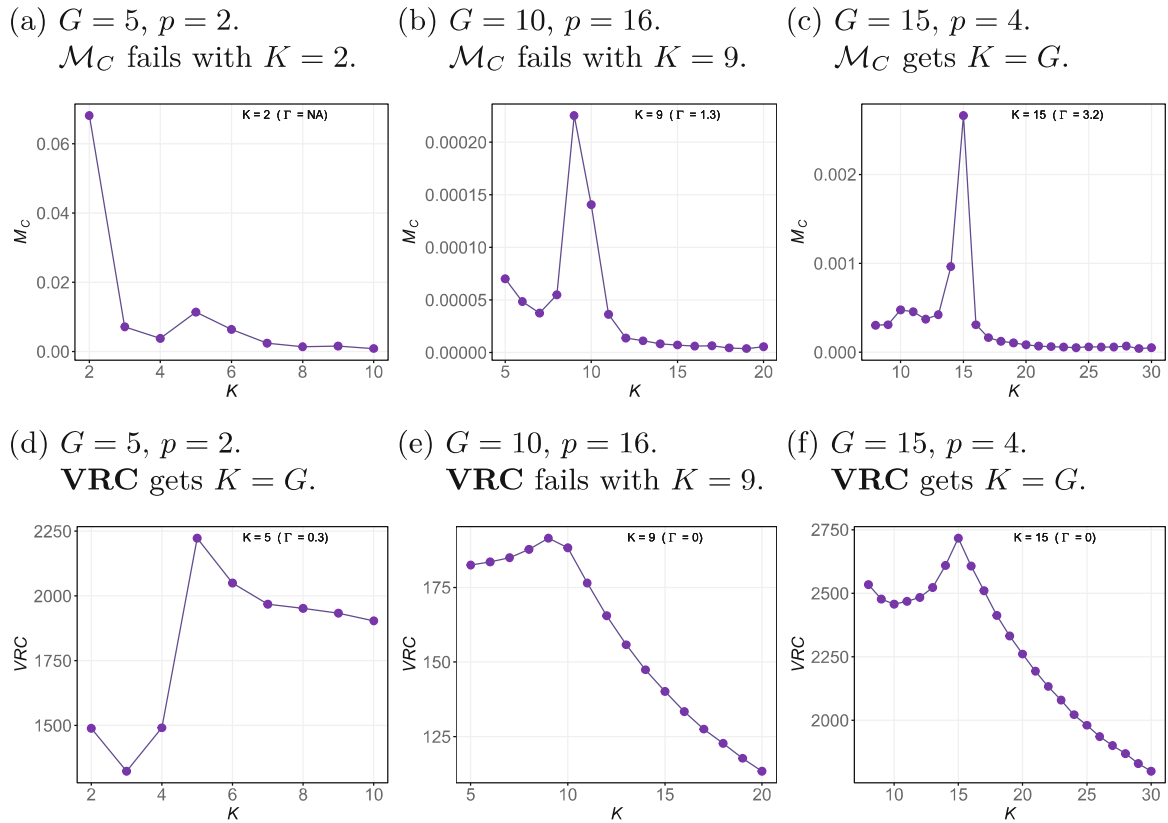


Fig. 1. Examples of \mathcal{M}_C (top row) and \mathbf{VRC} (bottom row) for different K , when realistic clusters ($J^* = 0.01$) are used.

A comparison of the shapes of \mathcal{M}_C and \mathbf{VRC} can be found in Fig. 1, where we see that \mathcal{M}_C typically has a sharper peak. In the real data, \mathcal{M}_C seems to perform even better in detecting the right number of classes.

Table 1 shows that \mathcal{M}_C performs better than χ_R^2 , both in terms of the accuracy and the sharpness of the peak. We can therefore conclude that our extension of the χ^2 by incorporating the scores corresponding to ideal cluster frequency and within-standard deviation does add value to the metric.

5 Experiments on Real Data

We also perform our experiments in four different real data sets, where we detect the optimal number of clusters automatically using \mathcal{M}_C , and compare it to what is known about the data sets independently. In these experiments, we assume that the classes in the data sets form spherical clusters, easily separated by the K-Means algorithm, and, as a consequence, that the ideal number of clusters is the number of classes. If this assumption does not hold true for the data set under consideration, our metric will not work. Indeed, the definition of our metrics would also be invalid in that case.

5.1 Variable Selection

Since we are testing our metrics 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 selecting the best variables based on purity, we will iterate over various values of K and expect to see a clear peak for the Covariant metric when plotted against various K values. Note that we will use the same “best” variables for all other indexes to which we compare our Covariant metric. Therefore, there is no unfair advantage or biases in using the selected variables in favor of our metrics. The four data sets used are briefly described below.

5.2 Data Sets

Iris Data Set. The classic Iris data set [11] contains 150 flower measurements along four variables (*Sepal Length*, *Sepal Width*, *Petal Length* and *Petal Width*) from three different iris species (*Setosa*, *Versicolor* and *Virginica*). 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. We select the variables *Petal Length* and *Petal Width* as the variables (based on the highest purity) to use when looking for the best K .

We can now look at how the Covariant metric (\mathcal{M}_C) varies when we cluster with different K s. The dependence is shown in Fig. 2a. We can see that the ideal $K = 3$ clearly shows up as a peak in both distributions, much more clearly in \mathcal{M}_C .

Table 1. Accuracy and Γ of various metrics

Metric	Well-separated	Medium	Realistic
	($J^* = 0.34$)	($J^* = 0.21$)	($J^* = 0.01$)
\mathcal{M}_C	100.0% (28.6)	90.0% (18.0)	45.0% (7.0)
χ_R^2	45.0% (0.1)	65.0% (0.1)	35.0% (0.1)
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%
$f(K)$	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 $K = G$). Note that **Gap** and **$f(K)$** do not detect the ideal K using maximum or minimum, and therefore Γ is not reported.

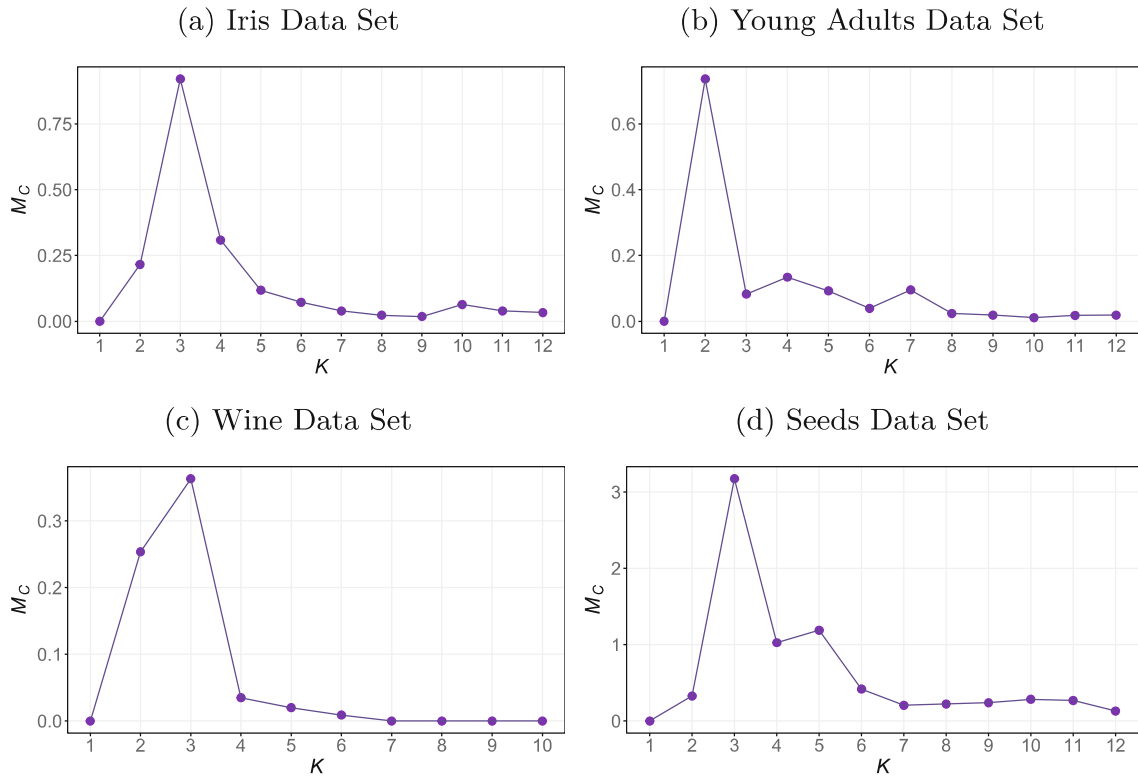


Fig. 2. Using our quality metrics to select the optimal number of clusters in various data set. The Covariant Metric \mathcal{M}_C shows clear peaks at the known number of clusters ($K = 3$ for the Iris, Wine and Seeds, and $K = 2$ for the Young Adults).

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%.

A blind K-Means clustering (with $K = 2$) using the four numeric variables is likely to segment the Young Adults data into male and female students. The ideal number of clusters is indeed two. In Fig. 2b, we have plotted \mathcal{M}_C as a function of K , and it shows a clear peak at $K = 2$.

The Wine Data Set. The publicly available Wine data set [1], from the UCI Machine Learning Repository [9], 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, we select the combination of *Alcohol*, *Ash*, *Flavanoids* and *OD280_OD315* based on the highest purity of 90.5%. The Wine data set also has three classes, and Fig. 2c shows that the Covariant metric (\mathcal{M}_C) has a clear peak at $K = 3$.

The Seeds Data Set. The publicly available Seeds data set [7] (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. From these combinations, we select *Area*, *Perimeter*, *Compactness* and *Asymmetry* based on the highest purity of 90.0% that we can get. KSelection-Seeds shows that the Covariant metric (\mathcal{M}_C) has a clear peak at $K = 3$, as expected.

6 Comparison with Other Indexes

Table 2. Performance comparison of our proposed metric and other indexes

Index	Data Set (Standardized)			
	Iris	YA	Wine	Seeds
G	3	2	3	3
\mathcal{M}_C	3 (1.15)	2 (65.92)	3 (2.32)	3 (13.30)
VRC	10 (0.01)	2 (0.27)	3 (0.14)	3 (0.10)
AIC	4 (0.04)	7 (0.02)	10 (0.00)	12 (—)
BIC	3 (0.24)	4 (0.07)	7 (0.01)	7 (0.01)
Sil. Wid.	2 (0.10)	2 (0.21)	3 (0.07)	2 (0.07)
DB	2 (0.38)	2 (0.12)	3 (0.08)	2 (0.01)
Gap	3	2	4	3
$f(K)$	2	2	2	2

Index	Data Set (Raw)			
	Iris	YA	Wine	Seeds
G	3	2	3	3
\mathcal{M}_C	3 (2.51)	2 (16.83)	3 (1.52)	3 (3.69)
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
$f(K)$	2	2	2	2

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 is reported between parentheses. It cannot be calculated at the end of the range $K = 12$.)

Some of the indexes to which we are comparing our metric may perform differently when the data set is standardized (such that all variables zero mean and unit standard deviation). For this reason, we study the performance of the indexes and our metric on both standardized data sets as well as the raw ones. Our proposed metric, however, does not require the data set to be standardized. In fact, since our metric takes into account the full covariance matrix on a per-cluster basis, it can be argued that it should perform as well or better in the *raw* data set.

We can see from Table 2 that our metric \mathcal{M}_C performs very well on standardized data sets, detecting the right K in all four data sets, while the other indexes seem to struggle. Of the seven other indexes considered, the **VRC** index seems to perform best with three right predictions. However, its significance measure (Γ) is low. When run on the data sets without any normalization, \mathcal{M}_C continues to perform well, as we can see in Table 2. The other indexes seem to perform marginally worse on the raw data sets than on the standardized ones.

Note that the sharpness of the peaks representing the right value of K , as measured by Γ is significantly higher for the Covariant metric \mathcal{M}_C , both in the standardized as well as the raw data sets (Table 2), when compared to any other index. The significance of the peaks (Γ) for our metric improves with standardization for three data sets, while decreases for the other one, which is consistent with our expectation that standardization should not affect its performance.

7 Limitations

The main motivation behind this work, in addition to pure academic interest, is to automate K-Means clustering such that it can be deployed in situations where automatic insight generation is desired. (For example, consider customer segmentation for marketing purposes where new customers are continually added to the database.) Since the impetus behind this work is automated processing, we have not attempted to prepare the data in any fashion.

The mathematical validity of the Covariant Metric (\mathcal{M}_C), being a ratio of two entities that may be thought of as χ^2 , is not yet fully established. It is similar to the odds ratio calculation commonly used in the data science community, but on shakier theoretical footing. It is hoped that other researchers may be able to find a more theoretically sound way of combining the components (defined in Eq. (4) and (9)) into a better metric than the one in Eq. (10). We can see from our results that there is information in the Covariant Metric when it comes to K selection (Fig. 2).

We may be able to use the significance of the peak, Γ as defined in Eq. (11), either directly or in combination with the peak value of \mathcal{M}_C in order to select the right K . We have not explored this idea further due to the uncertainty in the mathematical foundation of such an approach. Again, other researchers may be able to come up with theoretically defensible methods of using Γ .

Lastly, in defining our Covariant Metric, we implicitly assumed the need for a balanced data set (in which distinct classes occur with roughly the same

frequency, and with similar within-standard deviation), which may prove to be impractical in unattended deployments. While it is easy to see that the K-Means algorithm works best with balanced data sets, the usability of the Covariant Metric is limited to K-Means because of this assumption.

8 Conclusion

In this paper, we proposed a new quality metric for K-Means clustering and benchmarked it against existing indexes. From our comparative studies on synthetic data, we see that the Variance Ratio Criterion [5] (**VRC**) works remarkably well, followed closely by the Davies-Bouldin index [8] (**DB**). Our own index \mathcal{M}_C proposed in this article came in third when tested on synthetic data, but easily outperformed both **VRC** and **DB** in real data. Besides, the significance of the peak indicating the right K was substantially larger for \mathcal{M}_C .

All other indexes performed poorly on both synthetic as well as real data. Either \mathcal{M}_C or **VRC** seems to be preferable to the popular “elbow” method (which looks for a kink in the variation of the sum of squared errors, and is very subjective). Furthermore, our results indicate that both the Akaike and the Bayesian Information Criteria (**AIC** [3] and **BIC** [24]) are ineffectual in selecting the right K in K-Means clustering. The Gap Statistic [26] (**Gap**) performs slightly better than the information criteria, but it is prohibitively expensive, computationally.

Although more systematic exploration on more data sets is indicated, our Covariant Metric (\mathcal{M}_C) metric does show promise in the real data sets that we studied so far, as well as on an extensive collection of synthetic data. When it comes to discovering the right number of clusters, \mathcal{M}_C performed remarkably well. In fact, in real data, it outperformed the all other commonly used indexes of clustering quality by impressive margins.

Once we have a reliable metric for the quality of clustering, we can automate and build upon the current K-Means clustering algorithm. For instance, we can create scripts that will automatically select the optimal number of clusters (and possibly the best variables to use). Much like the forward selection or backward elimination processes in linear regression, K-Means clustering then becomes amenable to automatic optimizations. Furthermore, with robust metrics enabling automatic discovery of the right number of clusters, it may become possible to deploy K-Means clustering in situations where automated generation of insights without manual supervision is desired.

References

1. Aeberhard, S., Coomans, D., de Vel, O.: Comparison of classifiers in high dimensional settings. Technical report. 92-02, Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland (1992). [https://doi.org/10.1016/0031-3203\(94\)90145-7](https://doi.org/10.1016/0031-3203(94)90145-7)

2. Ahn, S., Fessler, J.A.: Standard errors of mean, variance, and standard deviation estimators (2003)
3. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974). <https://doi.org/10.1109/TAC.1974.1100705>
4. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. *Pattern Recogn.* **46**(1), 243–256 (2013). <https://doi.org/10.1016/j.patcog.2012.07.021>
5. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. *Commun. Stat.-Simul. Comput.* **3**, 1–27 (1974). <https://doi.org/10.1080/03610927408827101>
6. Campo, D., Stegmayer, G., Milone, D.: A new index for clustering validation with overlapped clusters. *Expert Syst. Appl.* **64**, 549–556 (2016). <https://doi.org/10.1016/j.eswa.2016.08.021>
7. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Żak, S.: Complete gradient clustering algorithm for features analysis of X-ray images. In: Piętko, E., Kawa, J. (eds.) *Information Technologies in Biomedicine*, vol. 69, pp. 15–24. Springer, Cham (2010). https://doi.org/10.1007/978-3-642-13105-9_2
8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979). <https://doi.org/10.1109/TPAMI.1979.4766909>
9. Dheeru, D., Taniskidou, E.K.: UCI machine learning repository (2017)
10. Feng, Y., Hamerly, G.: PG-means: learning the number of clusters in data. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) *Advances in Neural Information Processing Systems*, vol. 19. MIT Press (2006)
11. Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936). <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
12. Fränti, P., Rezaei, M., Zhao, Q.: Centroid index: cluster level similarity measure. *Pattern Recogn.* **47**(9), 3034–3045 (2014). <https://doi.org/10.1016/j.patcog.2014.03.017>
13. Hämmäläinen, J., Jauhiainen, S., Kärkkäinen, T.: Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms* **10**, 105 (2017). <https://doi.org/10.3390/a10030105>
14. Hamerly, G., Elkan, C.: Learning the K in K-means. In: *Advances in Neural Information Processing Systems*, vol. 17 (2004)
15. Hartigan, J.A.: *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley (1975)
16. Ishioka, T.: An expansion of X-means for automatically determining the optimal number of clusters. In: *Computational Intelligence* (2005)
17. Mahalanobis, P.C.: On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936)
18. Pakhira, M.: Finding number of clusters before finding clusters. *Procedia Technol.* **4**, 27–37 (2012). <https://doi.org/10.1016/j.protcy.2012.05.004>
19. Pelleg, D., Moore, A.W.: X-means: extending K-means with efficient estimation of the number of clusters. In: *ICML* (2000)
20. Pham, D., Dimov, S., Nguyen, C.: Selection of K in K-means clustering. *Proc. Inst. Mech. Eng. Part C-J. Mech. Eng. Sci.* **219**, 103–119 (2005). <https://doi.org/10.1243/095440605X8298>
21. Qiu, W., Joe, H.: Generation of random clusters with specified degree of separation. *J. Classif.* **23**(2), 315–334 (2006). <https://doi.org/10.1007/s00357-006-0018-y>
22. Qiu, W., Joe, H.: Separation index and partial membership for clustering. *Comput. Stat. Data Anal.* **50**, 585–603 (2006). <https://doi.org/10.1016/j.csda.2004.09.009>

23. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987). [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
24. Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978). <https://doi.org/10.1214/aos/1176344136>
25. Sieranoja, S.: How much K-means can be improved by using better initialization and repeats? *Pattern Recogn.* **93** (2019). <https://doi.org/10.1016/j.patcog.2019.04.014>
26. Tibshirani, R., Guenther, W., Trevor, H.: Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **63**, 411–423 (2002). <https://doi.org/10.1111/1467-9868.00293>
27. Yue, S., Wang, J., Wang, J., Bao, X.: A new validity index for evaluating the clustering results by partitional clustering algorithms. *Soft. Comput.* **20**(3), 1127–1138 (2015). <https://doi.org/10.1007/s00500-014-1577-1>
28. Zhao, Q., Fränti, P.: WB-index: a sum-of-squares based index for cluster validity. *Data Knowl. Eng.* **92**, 77–89 (2014). <https://doi.org/10.1016/j.datak.2014.07.008>