Cluster Metric Sensitivity to Irrelevant Features

Miles McCrory
Data Science Department
National Physical Laboratory
Hampton Road, Teddington, United Kingdom
m.mccrory@npl.co.uk

Spencer A. Thomas
Data Science Department
National Physical Laboratory
Hampton Road, Teddington, United Kingdom
spencer.thomas@npl.co.uk

***Abstract*—Clustering algorithms are used extensively in data analysis for data exploration and discovery. Technological advancements lead to continually growth of data in terms of volume, dimensionality and complexity. This provides great opportunities in data analytics as the data can be interrogated for many different purposes. This however leads challenges, such as identification of relevant features for a given task. In supervised tasks, one can utilise a number of methods to optimise the input features for the task objective (e.g. classification accuracy). In unsupervised problems, such tools are not readily available, in part due to an inability to quantify feature relevance in unlabeled tasks. In this paper, we investigate the sensitivity of clustering performance noisy uncorrelated variables iteratively added to baseline datasets with well defined clusters. The clustering quality is evaluated using labeled and unlabelled metrics, covering a range of dimensionalities in the baseline data to understand the impact of irrelevant features on clustering popular metrics. We show how different types of irrelevant variables can impact the outcome of a clustering result from k-means in different ways. We observe a resilience to very high proportions of irrelevant features for adjusted rand index (ARI) and normalised mutual information (NMI) when the irrelevant features are Gaussian distributed. For Uniformly distributed irrelevant features, we notice the resilience of ARI and NMI is dependent on the dimensionality of the data and exhibits tipping points between high scores and near zero. Our results show that the Silhouette Coefficient and the Davies-Bouldin score are the most sensitive to irrelevant added features exhibiting large changes in score for comparably low proportions of irrelevant features regardless of underlying distribution or data scaling. As such the Silhouette Coefficient and the Davies-Bouldin score are good candidates for optimising feature selection in unsupervised clustering tasks. Finally, we observe that standardizing and mean centering the data prior to clustering removes the discrepancies between Gaussian and Uniformly distributed irrelevant features and in general reduces variability in metrics between repeated cluster runs.**

***Index Terms*—Unsupervised Feature Selection; Clustering Metrics; Irrelevant Features; Clustering Sensitivity; Clustering Evaluation; k-means; Noisy data; Clustering Uncertianty**

I. Introduction

Clustering is an important unsupervised machine learning method and can be applied in pattern recognition, image segmentation and data mining problems, [1]. Clustering algorithms group similar points together based on a measure of distance or similarity of the features within the points, such as the Euclidean distance between their values or attributes.

In many applications of clustering, the ground truth labels are not available and the algorithms are used in a data exploration methodology or pattern recognition. Therefore, in order to have confidence in the resulting groupings, we must understand the impact of input data on the clustering results. Datasets are increasing in volume, dimensionality and complexity, often with distinct possible clustering tasks of interest based on subsets of the data. As such, for a given task of interest, the input data can contain redundant or irrelevant variables for that specific task. Feature selection methods can help identify and remove irrelevant features and have been widely used in supervised learning tasks, [2], [3], [4], [5], [6], [7]. However, these require ground truth labels in order to assess the impact of removing a feature for the learning task, such as increased accuracy (classification) or reduced error (regression), [8]. For unsupervised tasks, such as clustering, these labels are not available, nor is there a clear objective to assess feature selection. Therefore it is important to understand how sensitive clustering metrics are to the inclusion of irrelevant features for a task when assessing assigned clusters, [9]. Understanding these sensitivities can potentially identify candidate metrics for performing unsupervised feature selection, by providing a suitable objective to optimise.

In practice, we are unlikely to know *a priori* if variables in the data are uncorrelated to a given task. Typically exploratory data analysis is used to identify patterns and key features in the data that are then used to inform feature selection and down stream analysis. However, many real word problems are highly complex and the relationship between a feature and the target of interest may be unknown. Furthermore, individual inputs alone may not correlate to the task or output, but there may exist a nonlinear combination of inputs that describes the task well. Similarly, random variables can negatively impact clustering as the random distances will mask any useful information in the data, [10]. From a practical perspective, we may not know which of the variables are or are not important, but we would like our evaluation metrics to reflect if the data contains non informative features. This at least provides confidence in our interpretation of the metric values as 'good' or 'bad'. Additionally, this can enable feature selection through optimisation of the metric, thus providing an automated way to remove redundant or irrelevant features in an unsupervised data-driven way.

There exists a range of clustering algorithms that can be used effectively for specific tasks and applications, [11]. k-means is a centroid based algorithm and is one of the most

popular methods as it is quick to run and easy to implement, [1], [12]. The k-means algorithm minimise the distance of points within clusters, the within-cluster sum of squares, but maximise the distance of points between clusters, the between-cluster sum of squares. The algorithm iterates through a two step process, firstly assigning each point to the nearest clusters, and then updating the calculated centroids, [13]. The k-means algorithm requires the user to state the number of clusters *a priori*. If this is not known, the number k of clusters can be estimated using the so called 'elbow plots', such as the inflection point of total sum squared distances to cluster centroids as a function of k, [14], maximising metrics such as the Silhouette Coefficient as a function of k, or directly using data-driven embedding, [15]. Density based clustering methods such as DBSCAN, [16] avoid the need to know the number of clusters *a priori*, however, they do require the selection of a minimum number of points and radius parameter. This replaces one unknown parameter with two, which additionally may not be easy or intuitive to optimise. Due to the cluster number being a comparatively simple and interpretable parameter to optimise, and the algorithms ubiquitous use in clustering problems, in this work we focus on k-means.

In this work we investigate the sensitivity of the k-means clustering algorithm to increasing levels of random variables in the input data in order to study the impact on clustering performance. This is effectively increasing the level of impact of the so called 'curse of dimensionality', [9]. We conduct experiments on datasets with ground truth labels to verify useful metrics for practical applications where such information is not available. By artificially adding increasing amounts of random variables to input data we can determine the impact of these irrelevant features relative to the informative features in the data. We define irrelevant here as features that are uncorrelated with the cluster label, modelled as randomly generated values, that are added across all cluster groups. Our results cover several datasets of different dimensions and we monitor the ratio of random variables to informative features, evaluating clustering performance using several metrics. We investigate Gaussian and uniformly generated random values, as well the effect of scaling the data. This work could easily be extended to other types of clustering algorithms and different distributions for sampling random numbers. To the best of our knowledge this has not been investigated previously with the literature surrounding clustering sensitivity examining the internal parameters of clustering algorithms, i.e ablation studies or (hyper) parameter tuning, [17], [18], [19], [20], or sensitivity analysis, [21], [22], [23].

II. Problem Formulation

A. Data

The datasets used in the experiments are called the Dim-sets datasets, [24]. There are four different datasets available each with a different number of dimensions, $D = (32, 64, 128, 256)$, referred to as Dim-D where D is the dimensionality. The datasets all have 1024 data points and 16 clusters each made up of 64 data points. These clusters are generated to have a Gaussian distribution with clusters that are well separated in all dimensions. All of these datasets have associated ground truths and the initial clustering metric scores for these dataset can be used as a baseline reference point for all other datasets when irrelevant features are iteratively added.

B. Evaluation Metrics

There are various metrics that can be used to measure the performance of clustering algorithms. Clustering metrics assess performance in two main ways; either by comparison of predicted labels to a ground truth or by measuring spatial distances within and between clusters. In this study we have the associated ground truth labels and evaluate the clustering results with this information.

Normalised Mutual Information (NMI) compares a clustering outcome (X) to the ground truth clustering labels (Y) defined as

$$\text{NMI} = \frac{\text{MI}\left(X, Y\right)}{\text{mean}\left(H\left(X\right), H\left(Y\right)\right)} \ , \tag{1}$$

where MI() is the mutual information, $H(z)$ is the entropy of z. This has an upper bound of 1 indicating perfect clustering assignment and a lower limit of zero for incorrect clustering results, [25].

The Rand Index (RI), [26] is a measure of similarity between two sets of data groupings, in our case this is the similarity between the cluster results and the ground truth labels and defined as

$$\text{RI} = \frac{a + b}{N} \tag{2}$$

where a and b are the number of true positives and true negatives respectively. N is the total number of points in the data. As we are investigating the impact of random variables on clustering results, we use the adjusted Rand Index (ARI)

$$\text{ARI} = \frac{\text{RI} - E\left[\text{RI}\right]}{\max(RI) - E\left[\text{RI}\right]} \ , \tag{3}$$

where $E\left[\ \right]$ is the expectation value due to the random elements in the k-means initialisation leading to differing a and b across runs. This formulation ensures that random labels will have scores near zero. This has an upper bound of 1 indicating perfect clustering assignment.

The Silhouette Coefficient (S) measures the clustering results assuming the desired outcome is dense and separated clusters. It is defined using the mean distance between a point and all points within the same cluster (d_w) and the mean distance between a point and all other points in the nearest cluster (d_n)

$$\text{S} = \frac{d_n - d_w}{\max\left(d_n, d_w\right)} \ . \tag{4}$$

This has a lower limit of -1, sparse and overlapping clusters, and 1 for dense well separated clustering.

The Davies-Bouldin Index (DB), [27] compares the similarity of each cluster with the next most similar cluster within the dataset, averaged over all k clusters. The DB is calculated

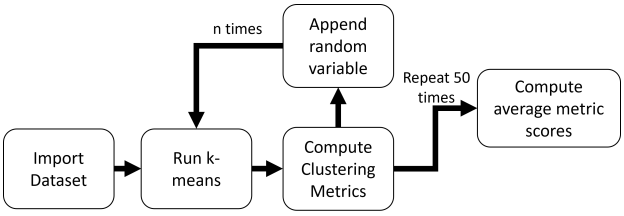

Fig. 1. Workflow for the experimentation used in this work.

using the average distance of all points to the centroid (δ) within each cluster, and the distance between centroids for pairs of clusters (Δ_{ij}),

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\delta_i+\delta_j}{\Delta_{ij}}\right) \tag{5}$$

The perfect Davies-Bouldin score is 0 which means clusters are well separated, clearly defined and dense. There is no upper bound for the Davies-Bouldin metric but a higher scores means clusters are poorly defined and overlap.

C. Experiments

To assess the impact of random variables on clustering performance we use a set of well defined clusters with associated ground truth labels outlined in Section II-A as our baseline. For each dataset we iteratively append one random variable to each instance in the dataset, increasing its dimensionality by one each time. Within the iteration cycle, we perform k-means clustering on the data and evaluate the results with the metrics outlined in Section II-B. The experimental workflow is summarised in Fig. 1. We then obtain the distribution of each metrics as a function of additional random variables. We represent the additional variables as a ratio of random variables to 'real' features in the data for comparison across different dataset dimensions. This will highlight the dependence of clustering performance on the proportion of random variables in the data.

Since we have the ground truth we know how many clusters there are in the datasets. Moreover, as the clusters are well defined and separated, we can compare clustering performance to this baseline, and therefore attribute any difference directly to the inclusion of the random variables. As we are iteratively increasing the dimensionality by adding random variables, and have several baseline datasets of increasing dimensionality of 'real' features (all of which well defined clusters), we are also able to examine general properties such as proportion of random numbers.

We generate random numbers using the mean (μ) and standard deviation (σ) of the features in the original data. We compare the effect of random numbers generated from a Gaussian distribution, R_G

$$R_G \sim \mathcal{N}(\mu_r,\, \sigma_r^2),$$

where

$$\mu_r = \mathrm{sign}(\mu+\sigma)\eta \ ,$$

$$\sigma_r = \sigma(1\,\mathrm{sign}\,\eta) \ .$$

Here η is a random number $[0, 1)$ and sign represents the sign ($\pm$) determined by an additional random number, $\mathcal{R} \in [0, 1)$,

$$\mathrm{sign} = \begin{cases} +, & \text{if } \mathcal{R} \geq 0.5 \\ -, & \text{otherwise} \end{cases} \ .$$

These are generated for each random value added hence, for the Gaussian distributed values, each random value added has a different mean (μ_r) and standard deviation (σ_r). We also consider random variables generated from a uniform distribution for identifying any impact due to noise distribution. For the uniform distributed random variable, R_U, we sample random numbers for the range $[-(\mu+2\sigma), +(\mu+2\sigma)]$

Finally, as scaling has been shown to affect clustering results, [10] we also examine these effects in our experiments. We compare unscaled data, generated with the distributions outlined above, with popular scaling methods. Specifically we consider *C*entered data, where each variable has the mean subtracted yielding a mean of zero in the scaled data, and *S*tandardized Centered data, where the variables are centered and scaling to have unit variance. Through the rest of the article we refer to these scaling as 'Centered' and 'Standardized' respectively.

We perform k-means with k specified from the ground truth labels in the data, 16 in our case. As our choice of implementation is the popular *k-means++*, [25]. We repeat each clustering experiments 50 times for each appended random variable to find an average clustering metrics and provide confidence bounds for these values. We have used Euclidean distance as our measure in all clustering experiments.

The proportion of added random variables is reported as a ratio of random variable to meaningful features in the baseline dataset. For instance, a ratio of 0:1 represents the baseline datasets in all cases, i.e. no added random variables, whereas a ratio of 2:1 means there are twice as many random variables in the data as there are informative features. In the case of a 2:1 ratio, Dim-32 has 64 random variables and 32 informative features, whereas Dim-128 has 256 random variables and 128 informative features. This allows comparisons to be made across the varying dimensionality datasets and generalisation of the results.

III. Results

Figure 2 illustrates the results of this work and summarising the dependence of appended random values on clustering performance metric under different scaling methods and when using different distributions to sample the random numbers. For clarity we also look at the standard deviation of these curves in Fig. 3 which follows the same structure as Fig. 2.

Overall, we observe the same behaviour in centered and unscaled data for all metrics and both random number distributions. Each of the four dataset demonstrate the same dependence when unscaled or centered, indicating that dimensionality does not influence this. We also see comparable values and dependencies in the standard deviation plots in

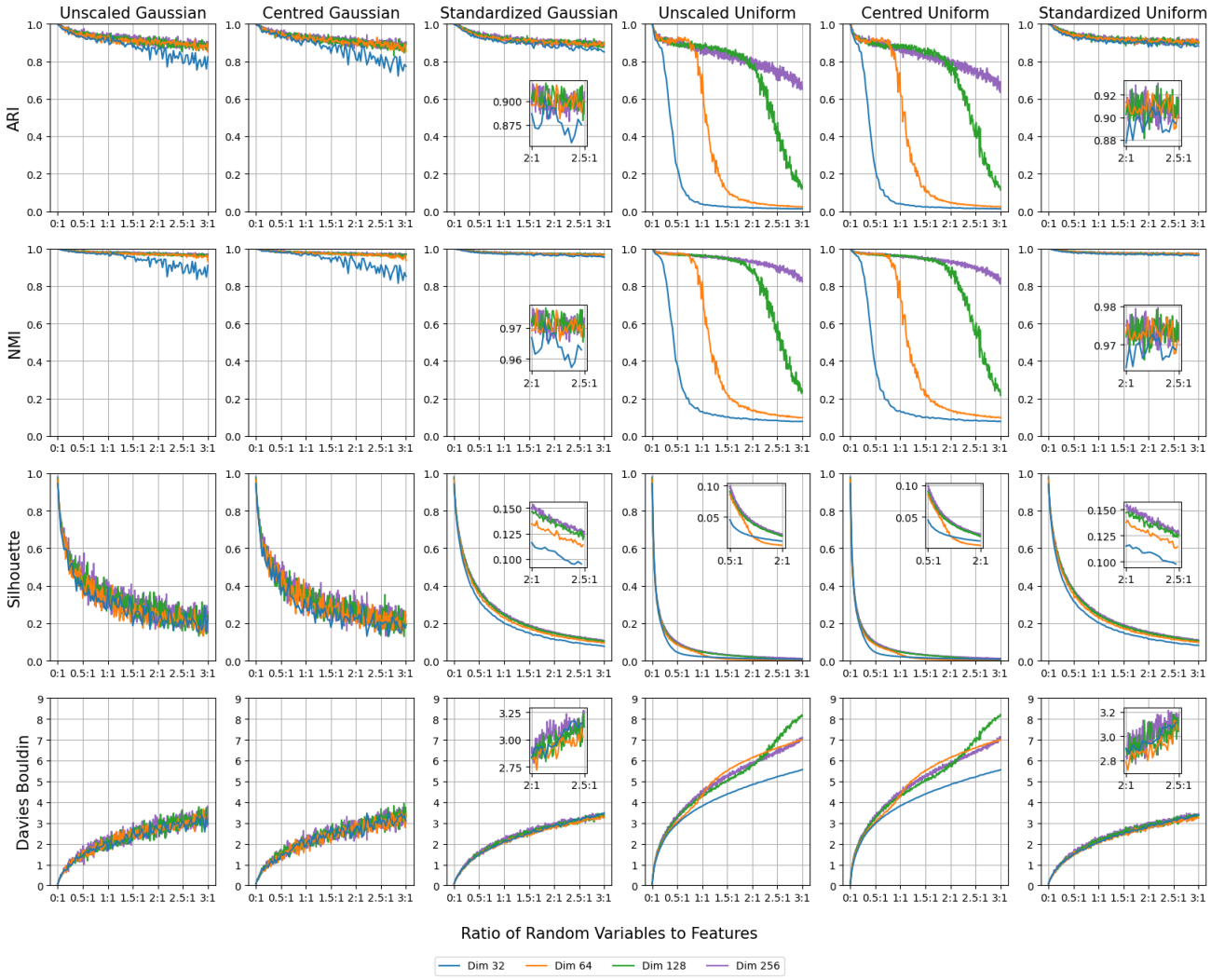

Fig. 2. A comparison of clustering performance with different random number generations and data scaling methods. Columns depict different scaling and random number generation methods, and rows illustrate clustering performance metrics averaged over 50 independent runs. The ratio of random variables to informative features ranges from 0:1 (baseline model) to 3:1 where 75% of the input data is randomly generated and therefore does not correlated to the cluster label. Note that higher Davis-Bouldin scores indicate worse clustering performance unlike the other metrics.

Fig. 3 for unscaled and centered data for both random number distributions. It is worth reiterating that the appended random variables generated from the Gaussian distribution each have a different mean and variance. As such the unscaled and centered data are different but appear to have the same dependency on appended random variables. Therefore, there is no observable benefit in centering the data when investigating random or uninformative features in the data for unsupervised tasks. For all metrics standardizing the data removes any discrepancy between random variables generated from a Gaussian or Uniform distribution. Moreover, standardizing the data reduces the standard deviation in performance scores in repeated runs for all configurations, with the exception of the Silhouette Coefficient for Uniform random variables which are comparable.

ARI and NMI behave qualitatively the same across all configurations in Fig. 2. When considering Gaussian random variables, scaling has little or no effect, with standardized data reducing the larger gradient of Dim-32 compared to the higher dimensional datasets making the dependence on appended random variables comparable across all dimensions. The larger gradient of Dim-32 is accompanied by increasing variation in ARI and NMI with increasing proportions of random variables, see Fig. 3. Both metrics are insensitive to large proportions of random variables, indicating high quality clustering performance even when 75 % of the data is random noise.

When considering uniform distributed random variables, scaled and centred data exhibit greater sensitivity to added random variables when assessed by ARI and NMI (noted by the steeper initial gradient in the curves in Fig. 2). However, there is a ‘tipping point’ where the scores rapidly decrease to near zero. The location of this tipping point

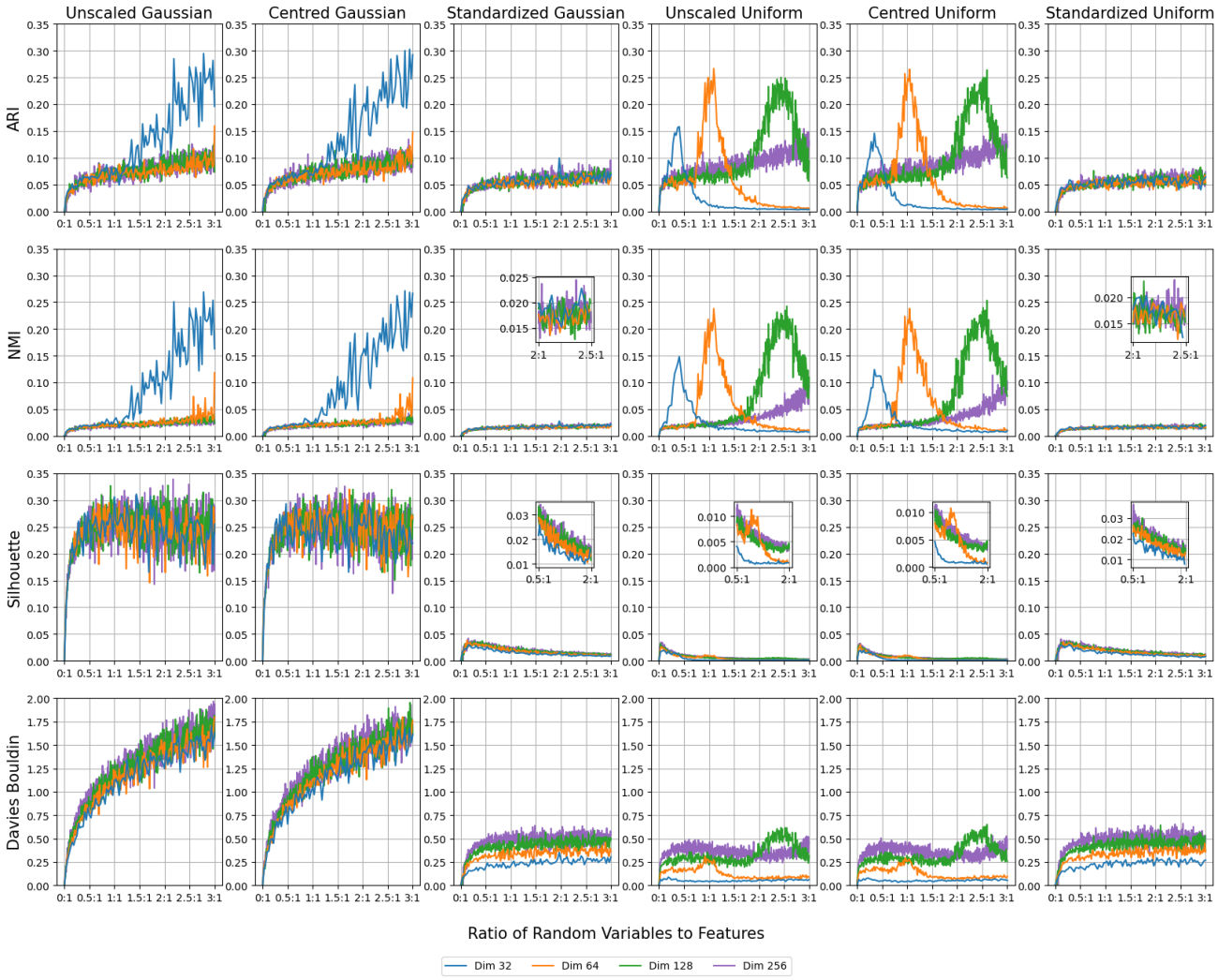

Fig. 3. Standard deviation (σ) value of clustering metrics score for 50 independent repeats. As in Fig. 2 we plot these values as a function of the proportion of random variables to features. Rows and columns are as in Fig. 2.

appears to be dependent on the dimensionality of the baseline data. Higher dimensions exhibit tipping points at higher proportions of random variables to features and the rate of reduction of score appears to be reduced. The start of these tipping points manifests as an abrupt increase in standard deviation (Fig. 3). The inflection point in Fig. 2 also corresponds with a maximum in the standard deviation in Fig. 3. Standardizing the data removes this dependency completely and resemble the standardized uniform curves closely resemble the standardized Gaussian curves in terms of scores (Fig. 2) and their variability (Fig. 3).

The Silhouette Coefficient and the Davies-Bouldin scores show clear dependence on the proportion of random variables, both showing larger gradients initially before reducing to a lower rate of change as observed in Fig. 2. The Silhouette Coefficient shows a rapid decrease in score from the baseline before indicating a plateau, this is more noticeable in the Uniform random number data. The Davies-Bouldin score has a comparatively lower initial gradient but appears to not to plateau, again more noticeable in the Uniform data. This may be due to the increase in intra-cluster distance with the addition of the random variables in line with the curse of dimensionality.

Specifically for the Gaussian random variables, standardizing data for both Silhouette Coefficient and Davis-Bouldin metrics reduces the variability seen in the curves in Fig. 2, see Fig. 3. Similarly to the scores, the standard deviation plateaus for the Silhouette Coefficient and continues to increase for Davis-Bouldin. When using standardized data, the Silhouette Coefficient yield scores at the lower end of the range seen in the unscaled and centered data, indicating lower quality clustering with increasing number of random variables added. Whereas, for the Davis-Bouldin metric, standardized data appear to have comparable values to the unscaled and centered data. For both metrics, standardized data drastically reduces

the variability between runs (Fig. 3), and in the case of the Davis-Bouldin the standard deviation plateaus.

For the Uniform random variables, both Silhouette Coefficient and Davis-Bouldin show rapid degradation of score with increasing proportions of random variables. This dependence is stronger, and exhibit much lower variation, than in the Gaussian case. For unscaled and centered data, both metrics also show some subtle structure in the curves that is dependent on the dimensionality of the baseline data. This results in clustering performance being worse for lower dimensional data at certain proportions of random variables. For example 1.5:1 Dim-64 has a higher Davis-Bouldin score than Dim-128 and Dim-256, whereas at 2.5:1 the Dim-128 curve has overtaken Dim-64 and Dim-256. At 3:1 it appears that Dim-256 is higher than Dim-64 and may exhibit the same pattern, though this is outside of the range of our analysis. This dependence on dimensionality manifests as peaks in the standard deviation plots in Fig. 3, albeit much smaller peaks than seen in ARI and NMI.

The Silhouette Coefficient shows similar patterns but these are obfuscated by the dynamic range of these curves and are visible in the insets in Fig. 2 and Fig. 3. Standardizing the data removes this structure for both metrics and leads to curves resembling the standardize Gaussian data.

IV. Conclusion

These results indicate that the Silhouette Coefficient and the Davies-Bouldin score are the most sensitive to irrelevant features in all cases. The Silhouette Coefficient exhibits rapid decrease in value in response to comparatively low levels of added irrelevant features indicating it is the most sensitive from a *perfect* feature set baseline. The Davies-Bouldin score also exhibits a rapid increase when irrelevant features are added, and its trend implies that it will not plateau. Both the Silhouette Coefficient and Davies-Bouldin score provide useful measurements of cluster quality that are sensitive to the addition of irrelevant features. As these metrics do not require ground truth labels, they are well suited as objective functions to optimise in feature selection for unsupervised tasks with unknown amounts of irrelevant features or when a *perfect* feature baseline is not available.

Conversely, ARI and NMI show a resilience to irrelevant features. For Uniform random numbers, this resilience is up to a critical point, that appears to be dependent on the dimensionality of the data. No critical points were observed with Gaussian random numbers and ARI and NMI maintained very high scores even at high proportions of irrelevant features relative to informative features in the data. This indicates that these metrics may not be useful for evaluating the clustering of noisy data, particularly if the noise is Gaussian distributed.

Finally, we also observe that Standardized data reduces the variability of the clustering results between runs and also provides comparable results between Gaussian and Uniformly distributed random variables. It also removes the appearance of tipping points in the Uniform random numbers that is dependent on the dimensionality of the baseline data.

References

[1] P. Berkhin. A Survey of Clustering Data Mining Techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer-Verlag, 2006.

[2] Nikolaos Papaioannou, Alkiviadis Tsimpiris, Christos Talagozis, Leonidas Fragidis, Athanasios Angeioplastis, Sotirios Tsakiridis, and Dimitrios Varsamis. Parallel Feature Subset Selection Wrappers Using k-means Classifier. *WSEAS Transactions on Information Science and Applications*, 20:76–86, 2023. Publisher: WSEAS.

[3] Suhang Wang, Jiliang Tang, and Huan Liu. Feature Selection. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 1–9. Springer US, Boston, MA, 2016.

[4] Nicholas Pudjihartono, Tayaza Fadason, Andreas W. Kempa-Liehr, and Justin M. O'Sullivan. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, 2:927312, June 2022.

[5] Jaime Zabalza, Jinchang Ren, Jiangbin Zheng, Huimin Zhao, Chunmei Qing, Zhijing Yang, Peijun Du, and Stephen Marshall. Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing*, 185:1 – 10, 2016.

[6] Zena M. Hira and Duncan F. Gillies. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, 2015:198363, 2015.

[7] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, October 2007.

[8] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1):131–156, January 1997.

[9] Michael E. Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? In Michael Gertz and Bertram Ludäscher, editors, *Scientific and Statistical Database Management*, Lecture Notes in Computer Science, pages 482–500, Berlin, Heidelberg, 2010. Springer.

[10] Leonard Kaufman, Peter J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. In *Finding Groups in Data: An Introduction to Cluster Analysis*, pages 1–67. John Wiley & Sons, Ltd, 1990.

[11] Dongdong Cheng, Qingsheng Zhu, Jinlong Huang, Quanwang Wu, and Lijun Yang. A local cores-based hierarchical clustering algorithm for data sets with complex structures. *Neural Computing and Applications*, 31:8051–8068, 2019.

[12] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, USA, January 2007. Society for Industrial and Applied Mathematics.

[13] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[14] Chunhui Yuan and Haitao Yang. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, 2(2):226–235, June 2019. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[15] Spencer Angus Thomas. Fast data driven estimation of cluster number in multiplex images using embedded density outliers. In *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8, 2022.

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[17] Dehua Peng, Zhipeng Gui, Dehe Wang, Yuncheng Ma, Zichen Huang, Yu Zhou, and Huayi Wu. Clustering by measuring local direction centrality for data with heterogeneous density and weak connectivity. *Nature communications*, 13(1):5455, 2022.

[18] Monika Krzak, Yordan Raykov, Alexis Boukouvalas, Luisa Cutillo, and Claudia Angelini. Benchmark and parameter sensitivity analysis of single-cell rna sequencing clustering methods. *Frontiers in genetics*, 10:1253, 2019.

[19] Pan Peng and Yuichi Yoshida. Average sensitivity of spectral clustering. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1132–1140, 2020.

[20] Sébastien Roux, Samuel Buis, François Lafolie, and Matieyendou Lamboni. Cluster-based GSA: Global sensitivity analysis of models with

temporal or spatial outputs using clustering. *Environmental Modelling & Software*, 140:105046, June 2021.

[21] Martin Kristiansen, Magnus Korpås, and Philipp Härtel. Sensitivity analysis of sampling and clustering techniques in expansion planning models. In *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*, pages 1–6. IEEE, 2017.

[22] Istvan Hajnal and Geert Loosveldt. The Sensitivity of Hierarchical Clustering Solutions to Irrelevant Variables. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 50(1):56, 1996.

[23] Peter O. Olukanmi and Bhekisipho Twala. Sensitivity analysis of an outlier-aware k-means clustering algorithm. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 68–73, November 2017.

[24] P. Fränti, O. Virmajoki, and V. Hautamäki. Fast agglomerative clustering using a k-nearest neighbor graph. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1875–1881, 2006.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.

[27] David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

CONTRIBUTION OF AUTHORS

Spencer Thomas designed the study and experiments. Miles McCrory and Spencer Thomas implemented and conducted all experiments. All authors interpreted the results and prepared the manuscript. The authors would like to thank Sam Bilson and Peter M Harris (NPL) for useful feedback on the manuscript.

SOURCES OF FUNDING

This work was funded by the Department for Science, Innovation and Technology through the National Measurement System.