Clustering Single-Cell RNA-seq Data: Impact of Data Binarization on Algorithmic Performance

Karolina Widzisz[1][a], Mateusz Kania[2][b], Joanna Zyla[3][c] and Andrzej Polański[1][d]
[1]*Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Akademicka 16, Gliwice, Poland*
[2]*Department of Applied Informatics, Silesian University of Technology, Akademicka 16, Gliwice, Poland*
[3]*Department of Data Science and Engineering, Silesian University of Technology, Akademicka 16, Gliwice, Poland*
{*karolina.widzisz, mateusz.kania, joanna.zyla, andrzej.polanski*}*@polsl.pl*

Keywords: scRNA-seq, Clustering Performance, Binary Data, Data Information Reduction.

Abstract: The primary objective of this study was to test the hypothesis that the binary information on the presence or absence of gene expression can sufficiently capture the inherent heterogeneity within single-cell RNA sequencing (scRNA-seq) data. This hypothesis posits that even without detailed expression levels, valuable insights about cellular diversity can be obtained. Utilizing this method can be particularly advantageous when analyzing large datasets, a common scenario in the field of scRNA-seq. In this paper, we evaluate clustering performance and cluster separability of a variety of model-based algorithms and distance-based methods to analyze both expression level data and threshold-encoded binarized data. We examined the performance of the Bernoulli-mixture model and Gaussian-mixture model. These were compared against traditional clustering techniques such as hierarchical clustering, K-means, and the Louvain algorithm on a range of scRNA-seq datasets. Our findings reveal that mixture models exhibit a lower dependence on the specific dataset compared to distance-based methods. Mixture models, particularly, demonstrate greater efficacy in accurately estimating the number of clusters present within the data. Among analyzed algorithms, the Bernoulli-mixture model stands out, outperforming distance-based approaches in several key aspects. Binary data, presence/absence of gene expression, seem to be indeed adequate to capture the heterogeneity of scRNA-seq data when clustering with methods specifically designed for binary datasets. The implications of this finding are significant, as it opens up new possibilities for simplifying data analysis in scRNA-seq studies without compromising the accuracy of the results.

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) generates extensive datasets measuring approximately 20,000 genes across thousands of cells, creating significant computational challenges in data analysis and visualization. With the increasing number of scRNA-seq experiments, understanding cellular heterogeneity in scRNA-seq data while managing data-specific issues remains challenging. scRNA-seq enables the identification of specific cell types and molecular targets for disease progression and drug response. Studying cell heterogeneity reveals subpopulations affecting disease pathology and drug resistance, enabling targeted personalized treatments.

Clustering algorithms should take into account specificity of scRNA-seq data. Identification of cellular subpopulations in scRNA-seq sequencing datasets presents challenges due to their large size and complexity, as well as occurrence of numerous dropouts in expressions of genes Zhang et al. (2023).

Existing research has extensively reviewed scRNA-seq clustering methods. In Kiselev et al. (2019) the authors described challenges in scRNA-seq clustering, including computational issues. They noted that large scRNA-seq datasets, with hundreds of thousands of cells, offer both challenges and opportunities. While large datasets may improve the power of analyses and the detection of rare cell types, they also make visualizing and interpreting clustering results difficult. Furthermore,

[a] https://orcid.org/0009-0003-2098-2232
[b] https://orcid.org/0000-0002-3605-0398
[c] https://orcid.org/0000-0002-2895-7969
[d] https://orcid.org/0000-0002-1793-9546

Widzisz, K., Kania, M., Zyla, J. and Polański, A.
Clustering Single-Cell RNA-seq Data: Impact of Data Binarization on Algorithmic Performance.
DOI: 10.5220/0013178800003911

In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 1*, pages 594-602
ISBN: 978-989-758-731-3; ISSN: 2184-4305

they discussed the issue of selecting the number of clusters, emphasizing that user-defined parameters significantly affect the clustering outcome. For some methods, such as k-means clustering, users explicitly specify the number of clusters, whereas for others, this number is determined indirectly through parameters such as the number of nearest neighbors in a graph.

In Petegrosso et al. (2019) the authors compare various clustering algorithms, including partition-based clustering (e.g. K-means, K-medoids), hierarchical clustering (HC), graph-based clustering (e.g. spectral clustering, clique detection, Louvain clustering), density-based clustering (e.g. DBSCAN, density peak clustering), neural networks (e.g. Kohonen networks), ensemble clustering, affinity propagation, and mixture models (e.g. Gaussian mixture models, hierarchical Dirichlet models). Using PBMC and breast cancer datasets, they found that current clustering methods work efficiently only with datasets of tens of thousands of cells. They emphasized the need to develop more scalable algorithms capable of handling larger datasets – up to a million cells. Similarly, in Duò et al. (2020) the authors compared 14 clustering methods based on dimensionality reduction techniques like PCA and t-SNE, and algorithms such as HC, K-medoids, K-means, ensemble clustering, nearest-neighbour graph clustering, density-based clustering, model-based clustering, and support vector machines (SVM), testing them on both simulated and real datasets.

Recently, it was hypothesized that transforming scRNA-seq data to binary format and therefore focusing on gene expression presence rather than level, can lead to improvements in the bioinformatics data analysis pipelines Bouland et al. (2021). Such an approach can lead to more robust and reliable results without the loss of sensitivity. In this paper, we further studied this hypothesis, by experimentally verifying whether threshold-encoding transformation in scRNA-seq data could capture dataset heterogeneity. We evaluated the performance of model-based and distance-based Bouveyron et al. (2019) clustering algorithms on the three scRNA-seq datasets, comparing their performance on original expression data and binarized data. Obtained results demonstrate that the gene expression presence alone is sufficient to capture genetic variability at the cellular level, potentially simplifying analysis of large scRNA-seq datasets.

2 METERIALS AND METHODS

2.1 Datasets

We utilized scRNA-Seq to analyze high-throughput molecular biology data, which featured sparse gene expression matrices containing thousands of genes across thousands of cells.

We selected three datasets from Chromium 10x platform scRNA-seq experiments, each with original group annotations for clustering evaluation. During preprocessing, we filtered out low-variance transcripts using Gaussian mixture model decomposition (using the threshold for the component with the lowest mean; Marczyk et al. (2019)), removed cells with fewer than 2,500 genes, and discarded transcripts lacking or having duplicate Ensembl IDs. Expression matrices were log-normalized using R *Seurat* package (v4.0; Hao et al. 2021).

The first dataset comes from a breast cancer (BC) study of circulating tumor cells Jordan et al. (2016) in women, obtained from the Gene Expression Omnibus database under accession number GSE75367. The second dataset contained raw scRNA-Seq data from RBC-depleted whole blood of COVID-19 patients and controls, published in Silvin et al. (2020). We filtered this dataset to include only COVID-19 patient samples from day 0, excluding the control group. For our third dataset, we used scRNA-Seq benchmark dataset of PBMC obtained from the Single Cell Portal of the Broad Institute Ding et al. (2020).

2.2 Data Preprocessing

For all datasets, we extracted genes with the highest variance, ranging from 5% to 50% of the largest variance across cells, in 5% increments. This created 10 subsets per dataset, each with varying matrix sparsity and information levels. Using variance-based gene selection minimizes the signal noise, reducing complications of expression level thresholding. We then applied binary coding as follows: 0 for non-expressed genes (expression = 0) and 1 for expressed genes (expression > 0). This represents a threshold-encoded approach for data binarization. Detailed characteristics of datasets are presented in Tab.1.

2.3 Examined Clustering Methods

In our study, we applied model-based clustering techniques and distance-based approaches to continuous (expression) and binary data representations of scRNA-Seq datasets.

Table 1: Summary of analyzed datasets.

Property	BC	COVID	PBMC
Genes	16,501	15,390	15,817
Genes per subset (5 - 50%)	825 - 8,250	770 - 7,695	791 - 7,908
Cells	232	2,564	3,222
Clusters	5	6	9

We utilized an independent multivariate Bernoulli mixture algorithm (BMM) for binary data Saeed et al. (2013) and a Gaussian mixture algorithm (GMM) Hennig et al. (2015) for continuous data, both using the Expectation-Maximization (EM) iterations McLachlan and Peel (2000) to identify model parameters. We assumed independence of components of multivariate distributions, which simplified computation and enhanced the algorithm's scalability. The EM algorithm was initialized with the K-means method Hartigan and Wong (1979), with the procedure repeated 20 times to find initial parameters with the highest likelihood.

For distance-based approaches, we employed Hierarchical Clustering Hubert and Arabie (1985), which organizes data hierarchically by treating each data point as a separate cluster and iteratively linking the nearest cluster pairs. We used Hamming distance for binary data and Euclidean distance for expression data, with Ward's linkage Ward (1963). We also examined the K-means algorithm, which partitions the dataset into K distinct clusters by maximizing within-cluster similarity and between-cluster distinctness. The algorithm assigns data points to clusters by minimizing the Within-Cluster Sum of Squares of distances to the cluster centroid. We repeated the centroid optimization procedure 20 times. Lastly, we applied the Louvain clustering method Blondel et al. (2008) for community detection in large networks. This method optimizes modularity by iteratively merging nodes into communities, then aggregating these communities into a new network until reaching optimal modularity. We constructed the network using Hamming distance matrices for binary data and Euclidean distance matrices for expression data.

Most of the time for real-life datasets the optimal number of clusters must be determined from the data. We tested algorithms with cluster numbers ranging from K = 2 to max (where max = 15 or the number of unique samples). The optimal model selection used the Bayesian Information Criterion (BIC) for BMM and GMM, and the Silhouette score Rousseeuw (1987) for other algorithms.

The analyses were performed using R (v4.2.2; R Core Team 2022). For HC, we used the *fastcluster* (v1.2.3; Müllner 2013), while for K-means clustering *stats* package. The Louvain method was implemented using the *igraph* package (v2.0.3; Csardi and Nepusz 2006). Additionally, we created implementations of the BMM and GMM.

2.4 Performance Evaluation

We conducted a comprehensive clustering evaluation using both clustering performance metrics and cluster separation metrics.

For clustering performance metrics, we used: (i) Adjusted Rand Index (ARI) Hubert and Arabie (1985), which measures cluster similarity by comparing sample pair assignments between predicted and ground truth clusters, offering reliable case-adjusted results; (ii) Fowlkes-Mallows Index (FMI) Campello (2007), which assesses performance through geometric mean precision and recall; (iii) Normalized Mutual Information (NMI) Fred and Jain (2005), which measures similarity between predicted clusters and ground truth labels—while Mutual Information (MI) tends to increase with cluster numbers, NMI reduces this bias by normalizing scores to 0–1, allowing for scale-invariant comparison; (iv) Error in estimated number of clusters (ENC), which measures how accurately the estimated cluster count matches the expected count, with values closer to zero indicating better estimation.

For cluster separation metrics, we applied: (i) Mean Silhouette Coefficient, which evaluates clustering quality by assessing how well data points fit their assigned clusters versus other clusters—higher values indicate more distinct clusters; (ii) Davies–Bouldin Index (DBI) Davies and Bouldin (1979), which measures average similarity between clusters based on within-cluster and between-cluster distances. Within-cluster distance represents the average distance from data points to their cluster centroid, while between-cluster distance measures the separation between centroids. A lower DBI indicates more distinct clusters, implying a more effective clustering solution.

We calculated each evaluation metric across all datasets and their combinations. To compare clustering method performance, the statistical inference using a t-test was performed, considering p-value$<$0.05 as statistically significant.

Additionally, to visualize how clustering results correspond to biologically labels, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) van der Maaten and Hinton (2008) utilized by Rtsne (version 0.17; Krijthe 2015).

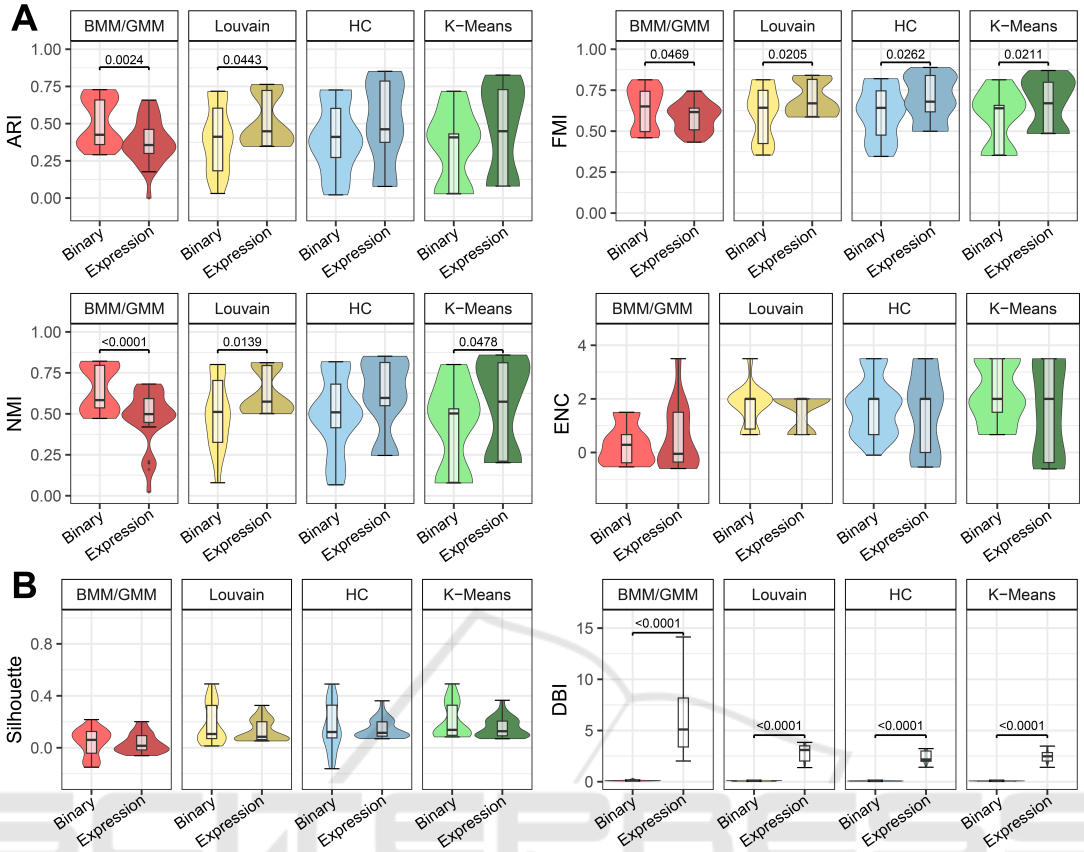

Figure 1: Evaluation of clustering on expression vs. Threshold-encoded datasets collectively presented on boxplots, measured by clustering performance metrics (A) and cluster separation metrics (B). Statistical significance was calculated with t-test.

3 RESULTS

3.1 Clustering Evaluation

We compared clustering results from expression and binary data across all datasets to evaluate differences between binary and continuous clustering methods (Fig.1).

Our analysis of clustering performance (Fig.1A) revealed statistically significant differences for model-based and Louvain clustering methods, except for ENC. K-means clustering showed significant differences in both FMI and NMI metrics, while HC demonstrated significant differences only in FMI. Expression data clustering generally yielded better overall performance, with model-based methods being the notable exception.

Regarding cluster separation measures (Fig.1B), we found no statistically significant differences in the Silhouette Index when comparing algorithms on threshold-encoded data versus expression data. However, the DBI showed significant differences favoring binarized data clustering, suggesting better cluster separation in this case.

BMM and GMM demonstrate consistent performance across different datasets, which is where these mixture models excel. A key advantage of mixture models is their ability to accurately estimate clusters. Their ENCs are typically the smallest, with an expected value of 0, and show better separability, particularly when measured by DBI. However, these algorithms may overlook subtle data variations, making them more suitable for global analysis than for identifying small, nuanced groups.

3.2 Dataset Effect

We observed that the results on specific datasets form clusters in Fig.1. Therefore, we compared the clustering performance of the algorithm pairs individually on each dataset, detailing the percentage of variance (Fig. 2).

Comparing BMM and GMM, the ARI shows BMM performs better on binary data across datasets,

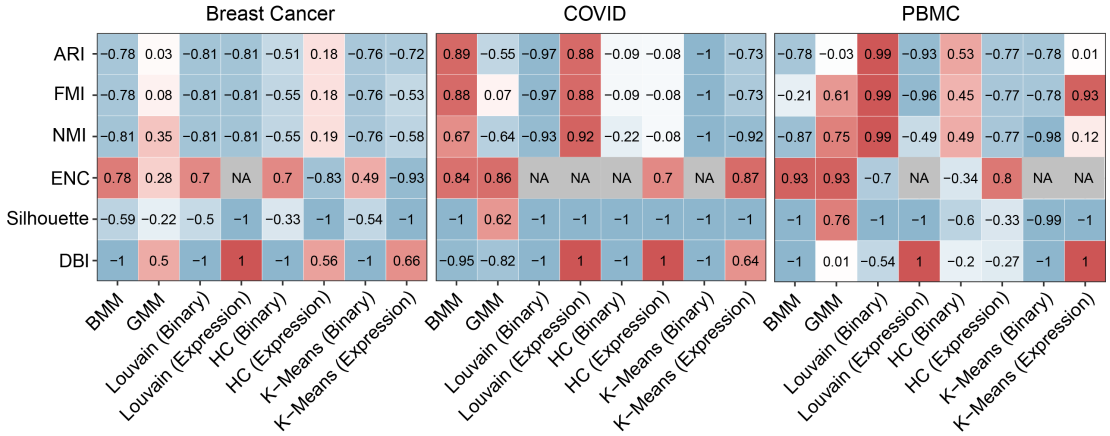

Figure 2: Spearman's rank correlation for the metrics considering subsetting due to the highest % of gene variance. "NA" appears if there are attributes with zero variance (with all elements equal).

with varying performance as variance increases: improving for COVID but declining for BC and PBMC. FMI follows similar trends, favoring binary data for COVID and BC, while expression data excels for PBMC. NMI consistently favors binary data across all datasets. ENC shows binary data is more stable, especially for PBMC, with COVID and BC showing comparable results between data types as variance increases. The Silhouette decreases with feature count in binary datasets (lowest for PBMC binary), while improving for expression data. DBI performs best with binary datasets, indicating better cluster separation.

For Louvain clustering, BC data shows comparable ARI between data types until 35% variance, after which binary data performance declines. COVID maintains similar ARI, with expression data showing slight advantages. PBMC binary data yields lower ARI than expression data, with FMI, NMI, and ENC following similar patterns. Silhouette is highest for binary COVID data and lowest for PBMC binary data. BC and PBMC show similar results across data types, while DBI indicates better cluster separation in binary datasets.

HC shows best ARI performance on BC expression data, with slightly lower results for binary data (5-40% variance). ARI worsens with increased variance in binary data. PBMC data shows poorest ARI, though binary data improves with variance while expression data declines. FMI and NMI mirror these patterns. ENC fluctuates significantly in PBMC binary data and performs worst in PBMC expression data, while remaining stable for COVID and BC binary datasets and performing best for BC expression data. The Silhouette is highest for COVID binary data, followed by COVID expression data, lowest for PBMC binary data, and similar across BC data types. DBI also performs better for binary datasets, indicating well-separated clusters.

K-means clustering yields worst ARI for PBMC and best for BC data, performing better overall on expression data. The trends for FMI, NMI, and ENC match those of ARI. The Silhouette reveals highest separability in COVID binary data, followed by COVID expression data, with poorest results in PBMC binary and BC expression data. The DBI also favors binary datasets, suggesting better separation of clusters.

Our analysis shows distinct trends in clustering algorithms across data types. BMM excels on binary data, particularly for COVID and BC datasets, but varies for PBMC data. Louvain performs well on binary data, with expression data better for COVID and PBMC datasets. HC is best for BC expression data and varies for PBMC data, while binary datasets generally give better DBI scores. K-means performs best on BC datasets and generally better on expression data, especially for ARI and Silhouette. Binary data often provides better cluster separation and stability, while expression data excels in specific contexts like COVID and BC datasets. The choice of algorithm and data type significantly impacts performance metrics, requiring careful consideration in clustering analysis.

3.3 BMM vs. Other Algorithms

We evaluated clustering algorithms on their preferred data types: BMM for binary data and others for continuous expression data (Fig.3). BMM outperformed GMM in ARI, FMI, and NMI, while matching other algorithms' performance. It excelled in estimating optimal cluster numbers (ENC) but scored lower on Silhouette Coefficient. For the DBI, BMM showed

superior results, with GMM being the least effective.

Overall, BMM demonstrates strong cluster separation and stability for binary data, despite some limitations compared to continuous expression data algorithms.

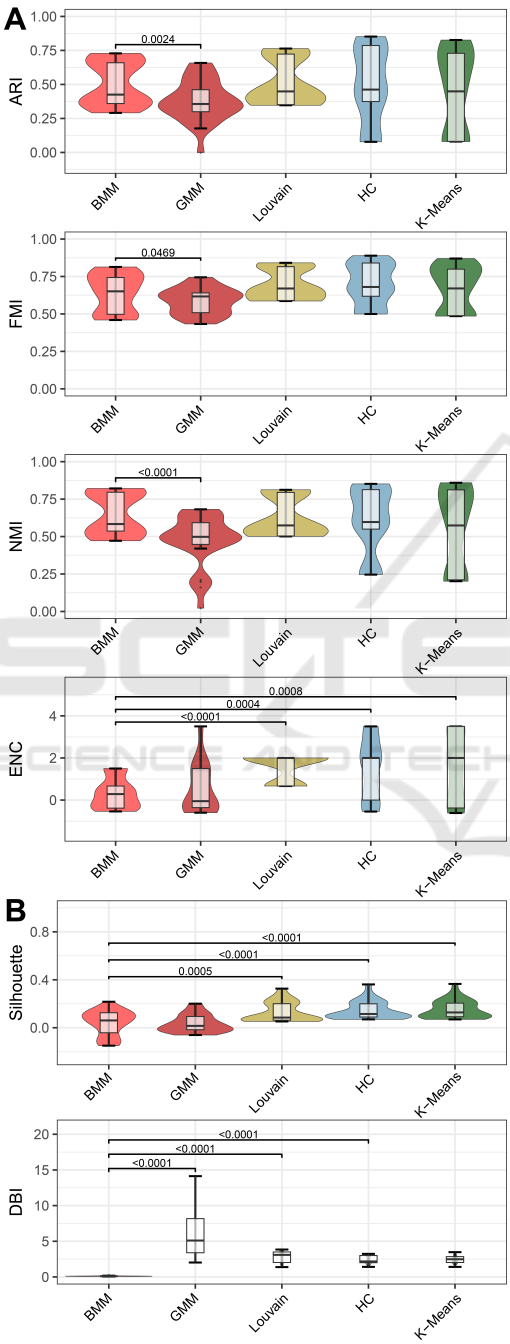

Figure 3: Evaluation of BMM clustering and algorithms used on expression data collectively presented on boxplots, measured by clustering performance metrics (A) and cluster separation metrics (B). Statistical significance was calculated with t-test.

3.4 t-SNE Visualization

We performed t-SNE dimensionality reduction on gene expression data, focusing on the top 25% of genes with the highest variability. This subset served as a strategic midpoint in our broader analysis, which examined gene subsets ranging from the top 5% to 50% of genes with the largest expression variance. By concentrating on the 25% range, we struck a balance between computational efficiency and capturing sufficient biological variation for meaningful clustering.

For visualization, we selected the COVID dataset, which contains labelled cell types including B-cells and T-cells, both known to comprise distinct subtypes. While the ground truth defined 6 clusters (Fig. 4A), Louvain found only 2 groups (Fig.4C), and both BMM and GMM detected 11 clusters (Fig.4B).

BMM provided better cluster separation, effectively identifying biological subtypes, especially within heterogeneous populations like B-cells and T-cells. Despite over-clustering, it produced more distinct and structured groupings of cell populations.

In contrast, GMM clusters showed poor definition and significant overlap. GMM struggled with expression data variability, producing clusters that lacked biological coherence and were less interpretable than BMM's results.

3.5 Computational Time

Our computational efficiency analysis across datasets showed significant variations by data type (Tab.2). HC achieved the fastest speeds consistently across all datasets and data types. K-means and model-based methods, took longer to process data, especially with the PBMC dataset. Binary data processing was generally faster, though K-means and HC performed slower with COVID and PBMC datasets than in case of expression data. Model-based methods benefited from threshold-encoding, which reduced data complexity and improved processing speed. While BMM shows better overall clustering performance, it remained slower than distance-based methods.

4 DISCUSSION

Our findings do not support the hypothesis that converting data from expression to binary worsens clustering; on the contrary, it is sufficient for capturing heterogeneity, as noted in Bouland et al. (2021). This indicates that the binarization of scRNA-seq data may not negatively impact the ability to identify di-

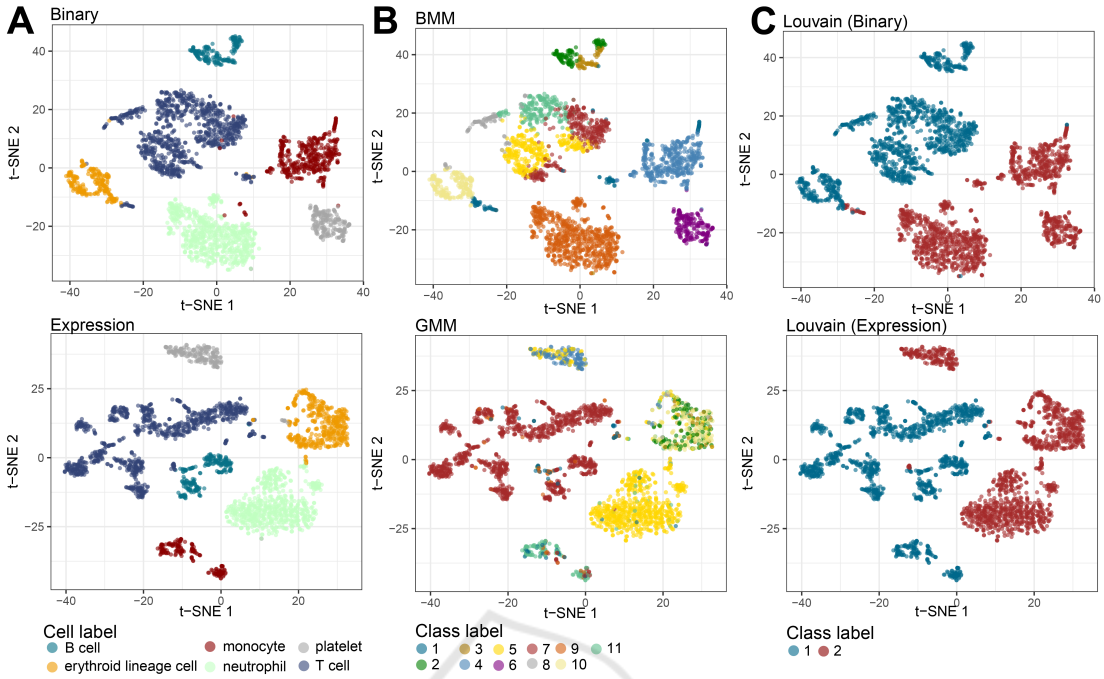

Figure 4: t-SNE projection of scRNA-seq for COVID dataset: original label of COVID set; binarized (up) and expression levels (down) (A), comparison of model-based algorithms (B), comparison of Louvain algorithm on binarized and expression levels data (C).

Table 2: Computational times for algorithms by data types and datasets (in seconds).

Data type		Binary			Expression		
Dataset		BC	COVID	PBMC	BC	COVID	PBMC
Algorithm	BMM	0.50	1.12	176.85	–	–	–
	GMM	–	–	–	7.69	543.58	1222.36
	HC	0.01	0.31	0.45	0.01	0.23	0.34
	K-means	0.66	12.03	29.79	1.50	9.62	16.33
	Louvain	0.01	1.27	3.38	0.02	4.34	3.80

verse patterns within the data. Bouland et al. (2021) focused mainly on binary differential analysis using logistic regression. In this paper, we evaluated a wider range of clustering algorithms, including mixture models, HC, K-means, and the Louvain algorithm. This allows for a more comprehensive understanding of the impact of data binarization on clustering performance.

Additionally, we investigated how the performance of clustering algorithms varies depending on the percentage of gene variance included in the analysis. It revealed that data binarization can be particularly beneficial when analyzing high-variance data.

Binarization is most effective in mixture models, which is why it is our primary recommendation. The BMM method's lower dependency on the dataset makes it the preferred choice in this category. ENC exhibits the smallest values, indicating good estimation of clusters, which is a huge advantage of the mixture models and the best cluster separability is observed for BMM. In this context, specifically designed for binary data mixture models show superior effectiveness, better handling the unique characteristics of binarized data for more accurate and reliable clustering. On binarized data, the BMM shows the same, if not better overall performance compared to distance-based methods. The binarization of scRNA-seq data for distance-based clustering algorithms may result in the loss of subtle expression level information, which can potentially impact the accuracy of cell type identification and differentiation. This is particularly significant in scenarios where minor variations in gene expression play a crucial role in distinguishing between closely related cell populations. Meanwhile, on continuous data, distance-based algorithms perform well. It is worth mentioning that BMM and GMM, in contrast to distance-based methods, appear to be less reliant on the particular dataset, highlighting the strengths of these mixture models. Nevertheless, it is important to consider the study's limitation: only 3 datasets were evaluated, and the comparison was restricted to metric-based and mixture approaches. Thus, future expansion with additional datasets to validate our observations is needed.

Computational time analysis revealed performance variations between binary and expression data types. BMM showed high efficiency with BC and COVID datasets in binarized form but slowed considerably with the larger PBMC dataset. GMM demonstrated increased computational demands with expression data across all datasets. In contrast, HC maintains rapid processing speeds across, making it attractive when time efficiency is crucial - though it may compromise accuracy in complex analyses. K-means exhibits higher computational demands for larger datasets, especially when processing binary data. While the Louvain algorithm remains efficient across most datasets and data types, its clustering performance decreases with larger binarized datasets. In summary, BMM performs best with smaller binary datasets but faces challenges with larger, more complex ones. HC and Louvain provide faster processing alternatives. Future development should prioritize improving BMM's scalability through parallel computing, as it remains the best-performing algorithm for threshold-encoded scRNA-seq data despite its computational limitations. In summary, BMM significantly outperforms traditional clustering techniques when applied to binary datasets.

T-SNE visualization demonstrated that BMM identified distinct subpopulations of T-cells and B-cells within the binarized COVID scRNA-seq data. The analysis revealed four T-cell subtypes, which likely correspond to CD8+ T-cells, CD4+ T-cells, regulatory T-cells, and memory T-cells. These subpopulations exhibit characteristic gene expression profiles that align with their known biological functions. For instance, CD8+ T-cells express cytotoxic genes such as GZMB and PRF1 Ramljak et al. (2021), while regulatory T-cells are marked by the expression of FOXP3 Dhawan et al. (2023) - established markers of functional and phenotypic diversity within T-cell populations. Furthermore, the analysis identified two B-cell subpopulations, which may represent plasma B-cells and memory B-cells. This distinction is supported by the differential expression of genes like PRDM1 Schultheiß et al. (2021) and CD27 García-Vega et al. (2024). Our findings, supported by existing literature, indicate that BMM's enhanced ability to differentiate cell types may be attributed to its sensitivity in detecting subtle gene expression variations that define these distinct immune cell subtypes.

An important biological conclusion is that, at least in some cases, the simple presence or absence of gene expression, rather than its level or intensity, might be sufficient for a meaningful analysis.

The field of data clustering offers a diverse range of methodologies for categorical data, which were beyond the scope of this study. Notable examples include K-Modes Huang (1998), Genetic K-Means Algorithm Krishna and Murty (1999), Maximum Dependency of Attributes Herawan et al. (2010), and Multiple Correspondence Analysis Xiong et al. (2009). Future work could explore these approaches, potentially revealing new insights and data structures in threshold-encoded scRNA-seq data.

ACKNOWLEDGEMENTS

This study was supported by the SUT grant for maintaining and developing research potential 02/090/BK_24/0043 [MK, AP], 02/090/BKM24/0045 [KW] and the Excellence Initiative - Research University program implemented at the Silesian University of Technology no. 02/070/SDU/10-21-01 [JZ].

REFERENCES

Blondel, V., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008.

Bouland, G. A., Mahfouz, A., and Reinders, M. J. T. (2021). Differential analysis of binarized single-cell RNA sequencing data captures biological variation. *NAR Genomics and Bioinformatics*, 3(4):lqab118.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Campello, R. J. G. B. (2007). A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recogn. Lett.*, 28(7):833–841.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Dhawan, M., Rabaan, A. A., Alwarthan, S., Alhajri, M., Halwani, M. A., et al. (2023). Regulatory t cells (tregs) and covid-19: Unveiling the mechanisms, and therapeutic potentialities with a special focus on long covid. *Vaccines*, 11:699.

Ding, J., Adiconis, X., Simmons, S., Kowalczyk, M., Hession, C., et al. (2020). Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature Biotechnology*, 38:1–10.

Duò, A., Robinson, M. D., and Soneson, C. (2020). A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7:1141.

Fred, A. L. and Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:835–850.

García-Vega, M., Llamas-Covarrubias, M. A., Loza, M., Reséndiz-Sandoval, M., Hinojosa-Trujillo, D., et al. (2024). Single-cell transcriptomic analysis of b cells reveals new insights into atypical memory b cells in covid-19. *Journal of Medical Virology*, 96.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., and et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13).

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28:100.

Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (2015). *Handbook of Cluster Analysis*. Informa.

Herawan, T., Deris, M. M., and Abawajy, J. H. (2010). A rough set approach for selecting clustering attribute. *Knowledge-Based Systems*, 23(3):220–231.

Huang, Z. (1998). Huang, z. extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Jordan, N. V., Bardia, A., Wittner, B. S., Benes, C., Ligorio, M., et al. (2016). Her2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature*, 537:102–106.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20:273–282.

Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.17.

Krishna, K. and Murty, M. N. (1999). Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439.

Marczyk, M., Jaksik, R., Polanski, A., and Polanska, J. (2019). Gamred – adaptive filtering of high-throughput biological data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc.

Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–18.

Petegrosso, R., Li, Z., and Kuang, R. (2019). Machine learning and statistical methods for clustering single-cell rna-sequencing data. *Briefings in Bioinformatics*.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramljak, D., Vukoja, M., Curlin, M., Vukojevic, K., Barbaric, M., et al. (2021). Early response of cd8+ t cells in covid-19 patients. *Journal of Personalized Medicine*, 11:1291.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Saeed, M., Javed, K., and Atique Babri, H. (2013). Machine learning using bernoulli mixture models: Clustering, rule extraction and dimensionality reduction. *Neurocomputing*, 119:366–374.

Schultheiß, C., Paschold, L., Willscher, E., Simnica, D., Wöstemeier, A., et al. (2021). Maturation trajectories and transcriptional landscape of plasmablasts and autoreactive b cells in covid-19. *iScience*, 24:103325–103325.

Silvin, A., Chapuis, N., Dunsmore, G., Goubet, A.-G., Dubuisson, A., et al. (2020). Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild covid-19. 182:1401–1418.e18.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Xiong, T., Wang, S., Mayers, A., and Monga, E. (2009). A new mca-based divisive hierarchical algorithm for clustering categorical data. In *2009 Ninth IEEE International Conference on Data Mining*, pages 1058–1063. IEEE.

Zhang, S., Li, X., Lin, J., Lin, Q., and Wong, K.-C. (2023). Review of single-cell rna-seq data clustering for cell type identification and characterization. *RNA*, page rna.078965.121.