p-ClustVal: A Novel p-adic Approach for Enhanced Clustering of High-Dimensional scRNASeq Data (Extended Abstract)

Parichit Sharma*, Sarthak Mishra*, Hasan Kurban*†, Mehmet Dalkilic*
*Computer Science, Luddy School of Informatics, Computing & Engineering
Indiana University, Bloomington, IN, USA
Email: parishar@iu.edu
†College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
Email: hkurban@hbku.edu.qa

***Abstract*—This paper introduces p-ClustVal, a novel data transformation technique inspired by p-adic number theory that significantly enhances cluster discernibility in genomics data, specifically Single Cell RNA Sequencing (scRNASeq). By leveraging p-adic-valuation, p-ClustVal integrates with and augments widely used clustering algorithms and dimension reduction techniques, amplifying their effectiveness in discovering meaningful structure from data. The transformation uses a data-centric heuristic to determine optimal parameters, without relying on ground truth labels, making it more user-friendly. p-ClustVal reduces overlap between clusters by employing alternate metric spaces inspired by p-adic-valuation, a significant shift from conventional methods. Our comprehensive evaluation spanning 30 experiments and over 1200 observations, shows that p-ClustVal improves performance in 91% of cases, and boosts the performance of classical and state of the art (SOTA) methods. This work contributes to data analytics and genomics by introducing a unique data transformation approach, enhancing downstream clustering algorithms, and providing empirical evidence of p-ClustVal's efficacy.**

***Index Terms*—p-Adic Numbers, Data-Centric AI, Single Cell RNA Sequencing, Unsupervised Learning**

I. Introduction

Clustering is a pivotal technique that organizes data into meaningful groups, unveiling inherent structures without prior labeling. It is especially critical in real-world applications, like biological data interpretation, where it helps in identifying distinct patterns and groups based on similarities among data points. In recent times, single cell RNA sequencing (scRNASeq) has marked a paradigm shift, enabling an in-depth exploration of data at an unprecedented single-cell resolution [1]. A primary challenge in scRNASeq data analysis is identifying distinct cell groups, akin to unsupervised clustering in machine learning [2]. Despite its potential, scRNASeq data processing is beset with challenges, primarily due to the noisy, low-resolution nature of the raw data, typically represented as an $m \times d$ matrix of cells (samples) and gene counts (features). The clustering efficacy is further impacted by data's high dimensionality, and sparsity [3].

This work introduces p-ClustVal, a novel data transformation technique for enhancing the clustering of scRNASeq data [1], [3], [4]. By operating in the transformed p-adic space [5], the method addresses challenges due to overlapping clusters, and amplifies the effectiveness of downstream clustering, enabling the discovery of previously obscured patterns and structures in the data. p-ClustVal is a user-friendly tool with minimal parameter tuning requirements, employing a unsupervised heuristic for optimal parameter selection. Our main **contributions** are, 1) a novel transformation technique (p-ClustVal) for enhancing cluster separability, 2) a data-driven and unsupervised heuristic for learning parameters from data, 3) an intuitive demonstration on synthetic and real-world data, and 4) empirical validation of p-ClustVal's effectiveness in improving classical and state-of-the-art clustering tools.

II. Methods

We define the notation as follows: let $\mathbf{x} \in \mathbb{R}^d$ represent a d-dimensional vector over the reals. The dataset $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ comprises m such vectors. We denote the i^{th} element of $\mathbf{x}$ by x_i. **P-adic** numbers were first introduced by Kurt Hensel in 1897. P-adic numbers are an extension of rational numbers obtained by completing the rational numbers with respect to the p-adic metric. p-adic-valuation ($V_p(n)$) [5], [6] is a mathematical function that finds the highest power of a prime number p that divides a given non-zero rational number n. It is defined as follows:

$$V_p(n) = \max\{i \in \mathbb{N} : p^i \mid n\} \tag{1}$$

We introduce a parametric space discretization technique inspired by p-adic-valuation. We define a transformation function $\theta : \mathbb{R} \rightarrow Z$, parameterized by separation (`sep`) and cohesion (`coh`), for any scalar $x_i \in \mathbb{R}$. For simplicity, here we denote `coh` by c, and `sep` by s. Specifically, our transformation function discretizes data to enhance the separation between different classes while preserving the proximity of data points within the same class. This function is given by:

$$\theta_{c,s}(x_i) = \lfloor \log_c(x_i^s) \rfloor \tag{2}$$

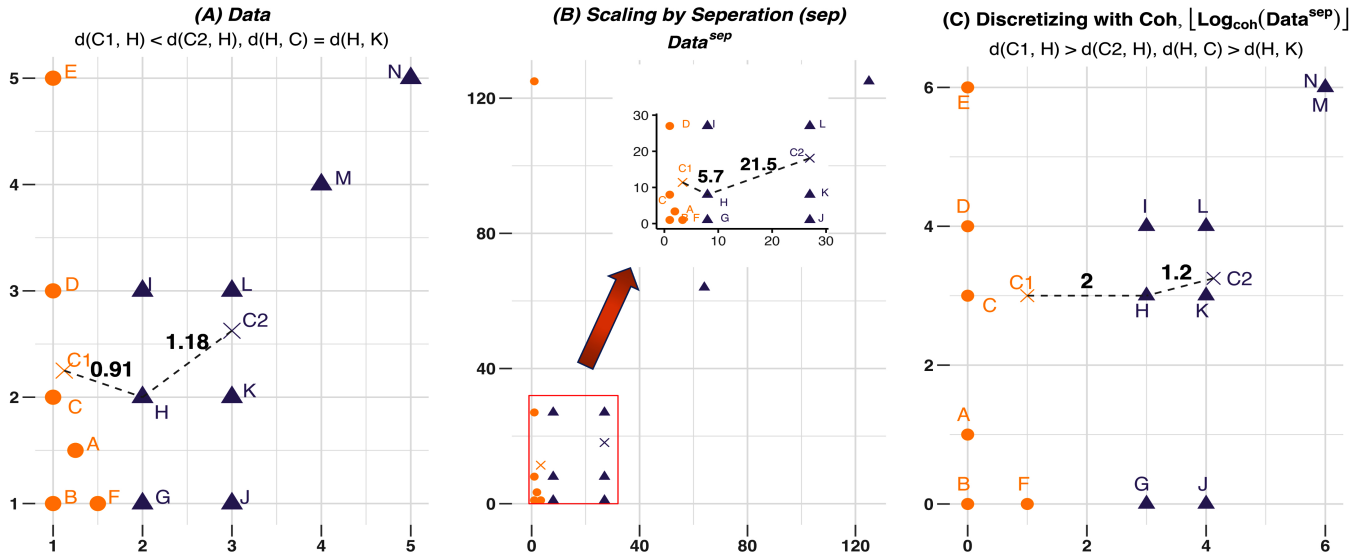

Fig. 1: Visualizing the effect of transformation on a simulated data. Orange and Violet points represent two different clusters. $C1$ and $C2$ are the respective centroids. (A): actual data, (B): after scaling the data with `sep` = 3. Figure in the inset highlights the data within the red region. (C): after discretizing with `coh` = 2. Dashed lines indicate the distance between cluster centers $C1$, $C2$ and H. Scaling pushes data points apart, but conserves the global spatial structure of data, as seen in plot-(B). Discretizing with `coh` brings data points in the same cluster closer, thus increasing the separation between different clusters.

Illustrative examples Fig. 1 illustrates the functioning of p-ClustVal. Points `G`, `H`, and `I` are equidistant from members of both clusters. Moreover, `H` is closer to `C1` than `C2`, and if the points are assigned based on their proximity to the centers then `H` will be misclassified as part of `C1`. Post transformation (Fig. 1-(C)), boundary points `G`, `H`, and `I` are relocated closer to their actual center, thereby improving the chances of correct clustering. The visualization on a real-world data with known class labels is shown in Fig. 2. The overlap between clusters, for instance: `PP cell and A cell`, is noticeably reduced in transformed data. Specifically, data points in cluster `PP cell` have gotten closer while moving away from neighboring clusters. Compared to actual data, the transformation has noticeably improved the chances of clustering `A cell`, `PP cell`, `Stellate cell`, `Acinar cell` and `Ductal cell`.

Data-centric Search for Optimal Parameters We need a heuristic to dynamically find parameters in unsupervised manner, without access to the ground truth. A reasonable value of `coh` would be the smallest distance between any two points within the dataset. This is reasonable, as all the closest points must be mapped into the same corner of their respective hypercube. Since, `coh` ensures that closest points are mapped to the same corner, hence, we set `coh` as the average distance of the data point with it's k closest neighbors. This is defined in (3).

$$coh = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} d(x_i, x_j)}{m} \tag{3}$$

k denote the number of neighbors and $x_i \neq x_j$.

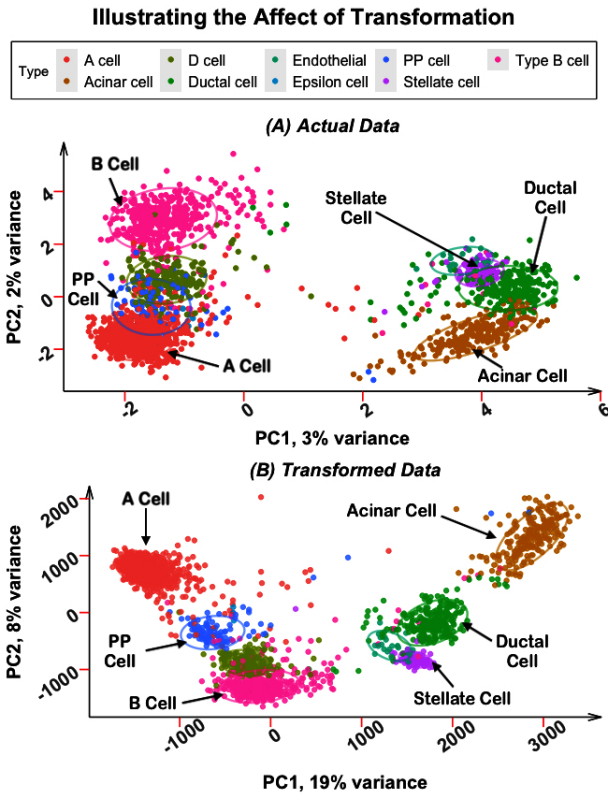

Fig. 2: Effect of p-ClustVal on Muraro data. Overlap between clusters has lessened driven by the reduction in spread of data, and separation between clusters have increased.

III. Experiments & Results

Experiments are conducted on 10 high dimensional and sparse scRNASeq datasets, varying in size, dimensions, cluster number, and amount of sparsity. Performance is measured via Adjusted Rand Index (ARI), that measure the extent of similarity between clustering labels and ground truth. Higher ARI indicate better alignment between ground truth and clustering labels. Performance is compared across three benchmarks.

*a) p-**ClustVal enhances classical clustering**:* The results are reported in Table I, and highlight the significance of p-ClustVal in enhancing clustering performance, as evidenced by the ARI scores. On majority of datasets, applying the transformation consistently improved the ARI of clustering algorithms, demonstrating the complementary nature of p-ClustVal.

*b) **Benchmarking with dimenionality reduction**:* Dimension reduction is commonly used either as a preprocessing step to filter noise or to obtain a lower dimensional representation of data before clustering [7]–[9]. Results (Table II) demonstrate that applying dimension reduction on transformed data reasonably improves the clustering accuracy.

*c) **Benchmarking SOTA clustering tools**:* We compare popular scRNASeq clustering packages [10]–[13]. Table III provide the results. Notably, no single method unanimously outperform others. On most datasets, p-ClustVal either improves the performance significantly or achieve similar performance as raw data.

TABLE I: Benchmarking classical clustering

Dataset	Adjusted Rand Index (ARI)							
	k-means		FZKM		EM		WAC	
	R	T	R	T	R	T	R	T
Pollen	0.60	**0.79**	0.09	**0.35**	0.28	**0.61**	0.76	**0.95**
Darmanis	0.52	**0.72**	**0.27**	0.25	0.24	**0.30**	0.49	**0.65**
Usoskin	0.09	**0.45**	0.10	**0.12**	0.07	**0.47**	0.02	**0.69**
Mouse Pan.	0.29	**0.56**	0.15	**0.54**	0.37	**0.45**	0.21	**0.48**
Muraro	0.56	**0.74**	0.40	**0.79**	0.36	**0.44**	0.62	**0.92**
Limb	0.14	**0.66**	0.05	**0.74**	0.10	**0.53**	0.008	**0.57**
Trachea	0.11	**0.58**	0.02	**0.59**	0.05	**0.43**	0.23	**0.45**
Lung	0.21	**0.49**	0.12	**0.71**	0.11	**0.53**	0.19	**0.42**
Diaphragm	0.25	**0.83**	0.09	**0.55**	0.15	**0.63**	0.25	**0.97**
Spleen	0.31	**0.48**	0.21	**0.27**	**0.41**	0.23	0.35	**0.77**

'R' and 'T' indicate raw and transformed data. **Bold** indicate better results.

TABLE II: Comparing the clustering with dimension reduction

Dataset	Adjusted Rand Index (ARI)					
	PCA		tSNE		UMAP	
	R	T	R	T	R	T
Pollen	0.66	**0.74**	0.73	**0.74**	0.64	**0.67**
Darmanis	0.57	**0.74**	0.50	**0.60**	0.44	**0.62**
Usoskin	0.12	**0.55**	0.06	**0.60**	0.04	**0.54**
Mouse pan.	0.21	**0.48**	0.31	**0.36**	0.36	**0.40**
Muraro	0.51	**0.83**	0.63	0.62	0.63	**0.76**
Limb	0.08	**0.62**	0.35	**0.51**	0.34	**0.51**
Trachea	0.12	**0.55**	0.31	**0.55**	**0.72**	0.51
Lung	0.20	**0.40**	30	**0.33**	0.31	**0.37**
Diaphragm	0.25	**0.81**	0.38	**0.53**	0.40	**0.62**
Spleen	0.30	**0.52**	0.25	0.25	**0.31**	0.28

Blue indicate either similar performance or when difference is less than 0.01.

TABLE III: Benchmarking SOTA scRNASeq clustering tools

Dataset	Adjusted Rand Index (ARI)									
	Seurat		RaceID		SIMLR		SC3		ADClust	
	R	T	R	T	R	T	R	T	R	T
Pollen	0.77	**0.8**	0.58	**0.82**	0.09	**0.5**	0.97	0.97	0.10	**0.68**
Darmanis	0.61	**0.66**	0.46	**0.51**	0.59	**0.84**	**0.95**	0.84	0.23	**0.84**
Usoskin	**0.64**	0.57	NA	NA	NA	NA	0.88	0.88	0.18	**0.27**
Mouse pan.	0.46	0.46	**0.31**	0.25	0.29	**0.47**	**0.41**	0.36	0.54	**0.58**
Muraro	0.93	0.93	**0.83**	0.78	NA	NA	0.94	0.93	0.39	**0.93**
Limb	0.66	0.67	0.5	**0.56**	0.24	**0.52**	0.64	**0.94**	0.04	**0.58**
Trachea	**0.95**	0.93	0.49	**0.54**	0.31	**0.51**	0.48	0.49	0.18	**0.90**
Lung	0.6	0.6	0.17	**0.41**	0.16	**0.76**	0.48	0.49	0.16	**0.59**
Diaphragm	0.98	0.98	0.13	**0.97**	0.31	**0.65**	0.99	0.99	0.02	**0.98**
Spleen	0.89	0.88	0.40	**0.44**	0.22	**0.5**	NA	NA	**0.63**	0.30

NA indicate method failed to run on the specific dataset.

IV. Conclusion

We introduced p-ClustVal, an innovative approach towards creating alternate representations of high dimensional genomics data. Drawing inspiration from the mathematical theory of p-adic numbers, p-ClustVal leverages metric spaces based on p-adic-valuation to enhance cluster separation and discernibility. p-ClustVal combined with an unsupervised heuristic for parameter estimation, is validated on downstream clustering, and dimension reduction methods, minimizes tuning requirements and shows remarkable adaptability across various datasets.

References

[1] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[2] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[3] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[4] Aaron TL Lun, Davis J McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5, 2016.

[5] Kurt Hensel. Über eine neue begründung der theorie der algebraischen zahlen. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 6:83–88, 1897.

[6] Fernando Q Gouvêa and Fernando Q Gouvêa. *p-adic Numbers*. Springer, 1997.

[7] Yanglan Gan, Ning Li, Guobing Zou, Yongchang Xin, and Jihong Guan. Identification of cancer subtypes from single-cell rna-seq data using a consensus clustering method. *BMC medical genomics*, 11:65–72, 2018.

[8] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[9] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):1–11, 2017.

[10] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

[11] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods*, 14(4):414–416, 2017.

[12] Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, Xiang Zhou, Xingjie Shi, and Jin Liu. Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic acids research*, 50(12):e72–e72, 2022.

[13] Yuansong Zeng, Zhuoyi Wei, Fengqi Zhong, Zixiang Pan, Yutong Lu, and Yuedong Yang. A parameter-free deep embedded clustering method for single-cell rna-seq data. *Briefings in Bioinformatics*, 23(5):bbac172, 2022.