Selecting a sparse set of important features from high-dimensional data to enhance clustering: an R implementation

Git Repo: https://github.com/wx-zhu/ClustVarSelect

Team Members :
Zhiyuan Yu(zyyu@umich.edu),
Wenxuan Zhu (zhuwx@umich.edu),
Adrian Con Garcia (acong@umich.edu)

Introduction

In high-dimensional data analysis, identifying a small subset of features that are most informative for distinguishing between groups or clusters is crucial. This problem becomes particularly important in genomic research, where datasets often contain a large number of variables (e.g., gene/protein expression levels) with a low signal-to-noise ratio. Efficient feature selection methods help reduce data complexity and improve the performance and interpretability of various downstream tasks such as clustering, regression, and classification.

This topic is particularly important for scRNA-seq datasets containing thousands of genes, as the correlation and distance between cells are dominated by noise before any processing. More specifically, scRNA-seq captures the expression levels of thousands of genes across individual cells and is frequently used to identify putative cell types after clustering, yet only a small subset of these genes contributes to specific biological outcomes[1]. Thus, identifying the correct subset of discriminating features is crucial not only for improving the accuracy of clustering but also for effectively identifying cell types and biological programs that could provide insights into novel therapies.

Unsupervised feature selection is technically challenging due to the absence of true labels, and previous studies addressed this issue either by certain ranking metrics (e.g., Laplacian score[2]) or by using certain optimization frameworks with lasso-like penalties (e.g., sparse PCA[3], sparse k-means[4]). However, these methods are not scalable to large datasets and sensitive to the choice of hyperparameter values, which are difficult to determine a priori, such as the number of clusters in the datasets.

To address these issues, a method called **Sparse Manifold Decomposition (SMD)** was recently developed[5]. SMD offers an interpretable framework for identifying features that best separate clusters in scRNA-seq data. Despite the existence of other tools for identifying highly variable genes, such as **SCTransform in Seurat**[6], SMD presents unique advantages: it ranks features based on their contribution to separability and is less dependent on the preprocessing steps required by alternative approaches. Nevertheless, the original implementation of SMD is constrained by its reliance on traditional k-means clustering, which can struggle with initial guesses and local minima. While k-means was initially used in SMD, we propose exploring alternative clustering algorithms, such as power k-means, spectral clustering, or DBSCAN, to improve its performance and applicability to complex biological datasets.

From this project, we expect to implement and improve the SMD algorithm in **R**, making it scalable, efficient, and compatible with other widely-used tools such as Seurat and Scanpy. Additionally, we aim to propose modifications to the algorithm to address its current limitations and improve its performance. Ultimately, our goal is to contribute to more efficient genomic data interpretations, supporting advancements in personalized medicine and therapeutic strategies.

Problem

In high-dimensional data analysis, particularly in genomic research, identifying a subset of informative features is crucial for reducing noise and improving clustering accuracy. scRNA-seq datasets, which capture thousands of gene expression levels per cell, often suffer from low signal-to-noise ratios, making feature selection essential for accurate cell type identification and downstream analysis. However, the lack of true labels in unsupervised settings presents a significant challenge.

Existing methods like sparse PCA[3], sparse k-means[4], and Laplacian scores[2] are sensitive to hyperparameters, computationally intensive, and struggle to scale to large datasets such as cell atlases. These methods also fail to handle the irregular biological structures commonly found in scRNA-seq data, limiting their applicability for identifying biologically relevant features.

The Sparse Manifold Decomposition (SMD) algorithm[5] offers a promising solution but is constrained by its reliance on traditional k-means clustering and its inability to integrate multiple datasets or handle batch effects. Additionally, the current Python implementation is neither scalable nor compatible with widely used tools like Seurat[6] and Scanpy[7], necessitating improvements to enhance its usability and performance.

Algorithms

The ClustVarSelect package implements an integrated approach to feature selection in high-dimensional biological data through two complementary methodological innovations. At its core, the package combines power k-means clustering with Bregman divergences for robust cluster identification, and sparse manifold decomposition (SMD) for discriminative feature selection. These methods work in concert to address the challenges of analyzing complex biological datasets, particularly single-cell RNA sequencing data. The power k-means algorithm provides a stable foundation for identifying underlying data structure, while SMD leverages this structure to identify biologically relevant features. This implementation builds upon recent advances in machine learning and bioinformatics, extending them with novel computational and statistical enhancements.

Power K-means with Bregman Divergence

Core Algorithm Description

The Power K-means clustering algorithm represents a significant advancement over traditional k-means by incorporating an adaptive annealing scheme through power means. This approach addresses fundamental limitations of classical k-means clustering, particularly its sensitivity to initialization and tendency to converge to poor local minima. While techniques like k-means++ have attempted to address initialization challenges, they often struggle as data dimensionality increases. The power k-means implementation in ClustVarSelect, through the `power_kmeans_bregman` function, extends these capabilities further by incorporating Bregman divergences for handling exponential family data distributions.

Unlike traditional k-means which employs hard assignments, power k-means utilizes soft assignments weighted by a power parameter s. In soft assignments, data points are associated with multiple clusters simultaneously, with their degree of association determined by a weight. This weight is influenced by s, allowing for a more flexible and nuanced representation of the data's relationship to the clusters. The initial power parameter s, typically set to -0.5, plays a crucial role in controlling the annealing process. This parameter, combined with the specified number of clusters k and various dimension reduction options (PCA, spectral clustering, or no reduction), creates a more flexible and robust clustering framework.

Optimization Process

The core optimization process demonstrates significant advantages over traditional k-means through its iterative scheme. Each iteration computes pairwise distances using specified Bregman divergences, including the standard Euclidean distance, KL divergence for probability distributions, Itakura-Saito divergence for scale-invariant comparisons, and logistic loss for binary-like data. These distances are then transformed through power mean operations, which create a smoother optimization landscape early in the clustering process when s is closer to zero, allowing for more effective exploration of the solution space. As s decreases, the assignments gradually become harder, eventually approaching the same objective as k-means but with superior optimization properties.

A distinctive feature is the dynamic adjustment of the power parameter s during optimization. Every two iterations, s is decreased according to specific rules: reduction by 0.2 if s > -1.0, or multiplication by the learning rate eta (default 1.05) if s > -120.0. This gradual reduction implements an annealing schedule that effectively avoids poor local minima, a common pitfall in traditional k-means.

Implementation Features

The implementation maintains computational efficiency through vectorized operations while handling complex calculations. For Bregman divergence computations, specialized functions ensure numerical stability, particularly for KL and Itakura-Saito divergences where data positivity is crucial. The algorithm monitors convergence through a threshold parameter, stopping when cluster assignments stabilize or reach the maximum iteration limit.

Another significant advantage over traditional k-means is power k-means' ability to handle non-spherical clusters. While traditional k-means assumes spherical cluster shapes due to its Euclidean distance metric, the power k-means framework, especially with Bregman divergences, adapts to various cluster geometries more effectively. This flexibility proves particularly valuable in high-dimensional biological data where cluster shapes may be complex.

The output provides comprehensive analysis capabilities through a structured PowerKmeans object containing cluster centers, assignments, and dimension-reduced data, supported by specialized print, summary, and plot methods. The implementation maintains theoretical guarantees, ensuring all iterates remain within the data's convex hull while minimizing within-cluster variance.

Following the mathematical framework of Xu & Lange[8] and incorporating Bregman divergences as proposed by Vellal et al. (2022), this implementation effectively handles various data distributions while maintaining computational efficiency through closed-form updates. While it introduces a modest computational overhead compared to traditional k-means, the superior clustering quality and robustness make it particularly well-suited for applications where clustering accuracy is paramount, especially in analyzing complex, high-dimensional biological data where traditional methods often fall short.

Sparse Manifold Decomposition (SMD) with Power K-means

Core Algorithm Description

The SMD algorithm implemented in ClustVarSelect represents a significant advancement in feature selection for high-dimensional biological data, particularly single-cell RNA sequencing data. Building upon the framework introduced by Melton and Ramanathan[5], our implementation combines sparse manifold decomposition with power k-means clustering and Bregman divergences to identify discriminative features effectively.

The algorithm begins by accepting a high-dimensional input matrix X and a parameter k_guess specifying the estimated number of clusters present in the data. A key innovation in our implementation is the integration of parallel processing capabilities, allowing the algorithm to distribute computational workload across multiple CPU cores. This enhancement is particularly valuable for large-scale analyses of biological datasets, where computational efficiency is crucial.

The core methodology proceeds through several carefully orchestrated stages. Initially, the input features are normalized to ensure comparable scales across different measurements. The algorithm then employs an ensemble approach, generating multiple cluster proposals through either agglomerative clustering ("agglo") or power k-means clustering ("kmeans"). The number of trials defaults to twice the number of features, while the subsampling parameter n_sub typically uses 80% of the data points for each proposal, striking a balance between computational efficiency and statistical robustness.

Feature Selection Approaches

For feature selection, SMD implements two distinct classification approaches. The entropy-based method utilizes decision trees through the `find_classifier_dims_entropy` function, while the maximum margin approach employs L1-penalized Support Vector Machines via `find_classifier_dims_maxmargin`. Both methods work to identify features that effectively discriminate between different clusters, but they approach the problem from different mathematical perspectives. The entropy-based method is particularly effective at capturing nonlinear relationships in the data, while the maximum margin approach excels at finding sparse, linear separating boundaries.

Implementation Features

A crucial aspect of the implementation is its handling of feature importance scores. When z_score is enabled (default behavior), the algorithm compares the observed feature importance distribution against a null distribution generated by randomly shuffling the data. This standardization process, implemented through parallel processing for efficiency, helps distinguish genuine discriminative features from those that might appear important by chance. The `shuffle_data` function carefully preserves the marginal distribution of each feature while breaking inter-feature relationships, providing a robust null model for significance assessment.

The algorithm incorporates sophisticated convergence monitoring through a threshold parameter that checks for stability in cluster assignments. This approach helps ensure that the identified features are robust and not artifacts of premature convergence. The implementation also includes careful input validation and type checking to ensure reliable operation across different data types and experimental conditions.

For the classification step, the `one_tree` function implements a single-level decision tree approach to identify the most discriminative feature between any pair of clusters. Similarly, the `one_plane` function employs L1-penalized SVM to find features that maximize the margin between cluster pairs. These complementary approaches allow the algorithm to capture different aspects of feature importance, particularly valuable in complex biological datasets where multiple types of discriminative patterns may be present.

The output is structured as a ClustVarSelect object, providing a comprehensive view of the feature selection results. This includes not only the raw feature importance scores but also detailed statistics and rankings that help researchers interpret and validate the results. The implementation includes specialized print and summary methods that facilitate easy access to key findings and statistics.

Implementation Advances

Following Melton and Ramanathan's framework, our implementation brings several significant innovations to advance feature selection in high-dimensional biological data analysis. The core enhancement incorporates power k-means clustering with Bregman divergences, enabling robust analysis of complex biological data distributions. This is complemented by a flexible input framework that supports multiple data formats commonly used in bioinformatics workflows, including regular matrices, sparse matrices (dgCMatrix), SingleCellExperiment objects, and Seurat objects in `power_kmeans_bregman` function. The implementation intelligently handles these different formats, automatically selecting appropriate data layers and preprocessing steps based on the input type and user preferences.

The algorithmic architecture implements sophisticated dimension reduction options, allowing researchers to choose between PCA, spectral clustering, or direct analysis without reduction. This flexibility is particularly valuable when working with different types of biological data that may require specific dimensional analysis approaches. The implementation supports various Bregman divergences (Euclidean, KL, Itakura-Saito, and logistic), enabling effective handling of different data distributions commonly encountered in biological research.

A major technical advancement is the integration of parallel processing capabilities optimized for large-scale datasets. The implementation automatically detects available computational resources and efficiently distributes workloads across multiple CPU cores, particularly beneficial for computationally intensive operations such as feature importance calculation and null distribution generation. This parallel framework carefully manages random seed generation to ensure reproducibility while maintaining computational efficiency.

The statistical robustness of the implementation is enhanced through comprehensive quality controls and sophisticated feature importance assessment methods. The standardization process utilizes efficient parallel computation of null distributions, offering both standardized (z-score) and raw importance scores. This methodological approach ensures reliable feature selection even in the challenging context of noisy, high-dimensional biological data, where distinguishing genuine signals from background variation is crucial.

These innovations collectively represent a significant advancement in feature selection methodology, particularly for single-cell RNA sequencing data analysis. The combination of sophisticated statistical approaches, efficient computational implementation, and careful consideration of biological data characteristics makes it a powerful tool for modern high-dimensional data analysis in computational biology.

The effectiveness of this integrated approach has been validated through extensive testing on both synthetic data and real biological datasets, as documented in the original publications. The package provides comprehensive S3 methods for visualizing and interpreting results, including methods for printing, summarizing, and plotting the clustering and feature selection outcomes. These tools make it easier for researchers to interpret and validate their results, supporting the broader goal of identifying biologically meaningful features in complex, high-dimensional data.

This implementation represents a significant advance in the field of feature selection for high-dimensional biological data. By combining robust clustering methods with sophisticated feature selection approaches and integrating them with modern bioinformatics tools, the package provides a powerful and flexible toolkit for researchers working with single-cell RNA sequencing data and other high-dimensional biological datasets.

Results

To test the correctness and efficiency of our implementation of the Bregman power K-means, we simulated some synthetic datasets with varying sample sizes, feature counts, cluster numbers, outlier proportions, and within-cluster noise levels (Fig 1). We then systematically evaluated and compared our implementation of the Bregman power K-means to the regular K-means. In our experiments, the Bregman power K-means outperformed the regular K-means, particularly when the cluster number, noise levels, and/or outlier proportions increase (Fig 2). Based on these results, our implementation of the Bregman K-means is more robust and accurate than the regular K-means.

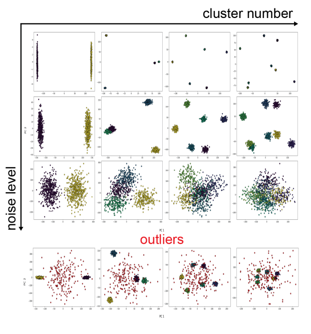

Figure 1. Examples of synthetic datasets

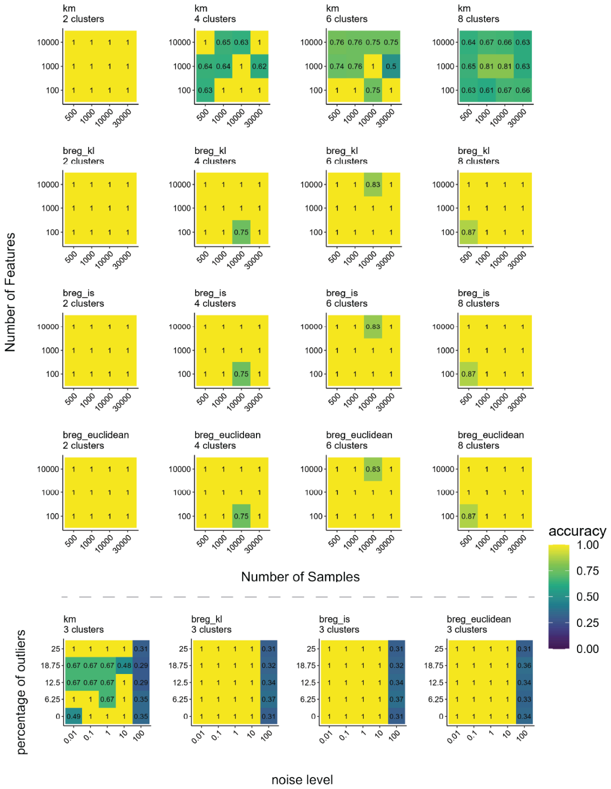

Figure 2. Comparing the accuracies of regular k-means (km) and Bregman power k-means variants with different divergence measures (Kullback–Leibler divergence (kl), Itakura-Saito divergence (is), and Bregman divergence using L2-norm as its convex function (euclidean)) on datasets with different sample sizes and feature numbers.

To test if our implementation of the Bregman power K-means is applicable to scRNA-seq data, we evaluated its performance on the gold-standard PBMC-3K dataset. Our implementation of the Bregman power K-means produced a clustering output more consistent with Seurat's output compared to the regular K-means algorithm, further supporting the superior performance of the Bregman K-means compared to the regular K-means (Fig. 3).

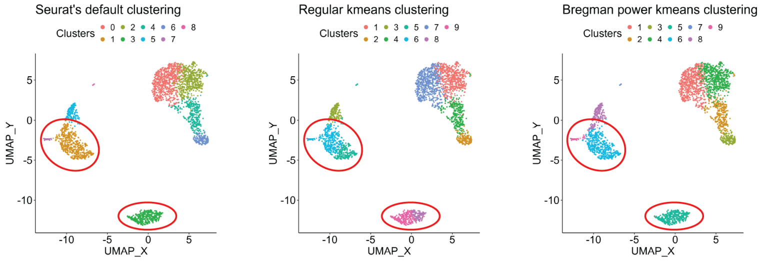

Figure 3. Comparison of clustering results between the Seurat, Bregment power K-means, and regular power K-means

To test our implementation of the SMD algorithm, we first argued that feature selection by SMD would improve the clustering quality, as SMD aims to increase the separability of potential clusters in data. In support of this hypothesis, applying Leiden clustering separately on feature sets selected by Seurat (based on normalized variance) and SMD (based on the relative contribution to the cluster separability) shows that SMD features improve the quality of the resulting clusters (Fig. 4).

We then argued that SMD selects features that not only improve cluster separability but are also more biologically meaningful, as features that improve cluster separability are more likely to be cluster markers. Consistent with our expectation, on the PBMC-3K dataset, our implementation of the SMD algorithm selects more marker genes than Seurat (Fig. 5). Since SMD selects more marker genes than Seurat on the PBMC-3K dataset, we speculated that pathway enrichment analysis on the SMD-only features (features that were selected by SMD but not Seurat) should output biologically meaningful pathways. Indeed, multiple immune pathways are enriched by the SMD-only features, which makes logical sense as the PBMC-3K dataset was collected from the peripheral blood.

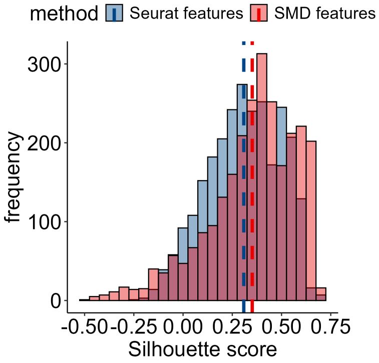

Figure 4. Clustering quality of Leiden clustering using Seurat or SMD's feature set.

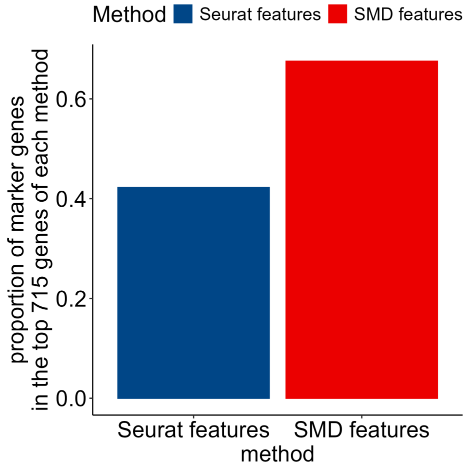

Figure 5. Proportion of marker genes in Seurat and SMD's top feature sets.

Finally, we tested our implementation of the Bregman power K-means on a monkey single-cell atlas (2 samples for each of the three developmental stages collected (CS8,9,11), and 56636 cells in total) that maps monkey gastrulation[1]. We set the cluster number to be equal to the actual number of clusters annotated in that dataset, and we used SCVI[4] to obtain the batch-free latent space. We then used the Euclidean version of the Bregman power k-means, since we didn't see significant differences between different divergence measures before, to cluster the dataset in the latent space. Our current implementation takes roughly 3 minutes on such a large dataset. Our Bregman power k-means performs reasonably well in identifying most of the major cell types (Fig 6). Although the mesoderm clusters identified by our Bregman power k-means are slightly different from the reference annotation, we think that it is reasonable because, at this stage of development, different mesodermal clusters are less distinguishable from each other.

However, when we attempted to run SMD on such a large dataset, it took hours to run and didn't significantly outperform Seurat, even after we parallelized the algorithm. While both the Bregman power K-means and SMD perform well on simple datasets, they still struggle to handle more complex atlas-level datasets. Overall, we conclude that our implementation of the Bregman power K-means and SMD behaves as expected, but these methods are more suitable for simple datasets.

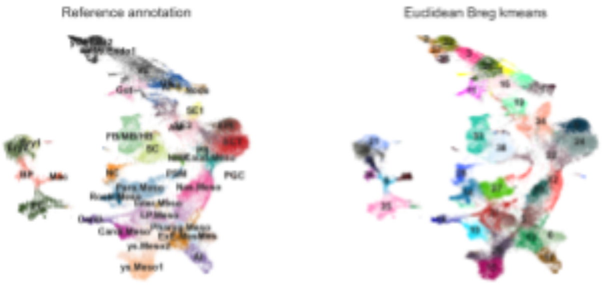

Figure 6. Applying the Euclidean version of the Bregman power k-means on the single-cell atlas of monkey gastruloids ADDIN EN.CITE ADDIN EN.CITE.DATA [1]. Left, reference annotation. Right, Bregman clusters.

References

1. Andrews, T.S., and Hemberg, M. (2018). Identifying cell populations with scRNASeq. Mol Aspects Med 59, 114-122. 10.1016/j.mam.2017.07.002.
2. He, X., Cai, D., and Niyogi, P. (2005). Laplacian score for feature selection. Advances in neural information processing systems 18.
3. Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. J Comput Graph Stat 15, 265-286. 10.1198/106186006x113430.
4. Witten, D.M., and Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. J Am Stat Assoc 105, 713-726. 10.1198/jasa.2010.tm09415.
5. Melton, S., and Ramanathan, S. (2021). Discovering a sparse set of pairwise discriminating features in high-dimensional data. Bioinformatics 37, 202-212. 10.1093/bioinformatics/btaa690.
6. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y.H., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell 177, 1888-+. 10.1016/j.cell.2019.05.031.
7. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19. 10.1186/s13059-017-1382-0.
8. Xu, J. and Lange, K. (2019). Power k-means clustering. In International Conference on Machine Learning, pp. 6921-6931.
9. Vellal, A., et al. (2022). Bregman Power k-Means for Clustering Exponential Family Data. ICML.