Multiple Mean Comparison for Clusters of Gene Expression Data through the t-SNE Plot and PCA Dimension Reduction

Yiwen Cao[1] and Jiajuan Liang[1,2,*]

[1]*Department of Statistics and Data Science, BNU-HKBU United International College, 2000 Jintong Road, Tang Jia Wan, Zhuhai 519087, China*

[2]*Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, 2000 Jintong Road, Tang Jia Wan, Zhuhai 519087, China*

Abstract: This paper introduces a novel methodology for multiple mean comparison of clusters identified in gene expression data through the t-distributed Stochastic Neighbor Embedding (t-SNE) plot, which is a powerful dimensionality re- duction technique for visualizing high-dimensional gene expression data. Our approach integrates the t-SNE visualization with rigorous statistical testing to validate the differences between identified clusters, bridging the gap between exploratory and confirmatory data analysis. We applied our methodology to two real-world gene expression datasets for which the t-SNE plots provided clear separation of clusters corresponding to different expression levels. Our findings underscore the value of combining the t-SNE visualization with multiple mean comparison in gene expression analysis. This integrated approach enhances the interpretability of complex data and provides a robust statistical framework for validating observed patterns. While the classical MANOVA method can be applied to the same multiple mean comparison, it requires a larger total sample size than the data dimension and mostly relies on an asymptotic null distribution. The proposed approach in this paper has broad applicability in the case of high dimension with small sample sizes and an exact null distribution of the test statistic.

Objective: Propose a two-step approach to analysis of gene expression data.

Gene expression data usually possess a complicated nonlinear structure that cannot be visualized under simple linear dimension reduction like the principal component analysis (PCA) method. We propose to employ the existing t-SNE approach to dimension reduction first so that clusters among data can be clearly visualized and then multiple mean comparison methods can be further employed to carry out statistical inference. We propose the PCA-type projected exact F-test for multiple mean comparison among the clusters. It is superior to the classical MANOVA method in the case of high dimension and relatively large number of clusters.

Results: Based on a simple Monte Carlo study on a comparison between the projected F-test and the classical MANOVA Wilks' Lambda-test and an illustration of two real datasets, we show that the projected F-test has better empirical power performance than the classical Wilks' Lambda-test. After applying the t-SNE plot to real gene expression data, one can visualize the clear cluster structure. The projected F-test further enhances the interpretability of the t-SNE plot, validating the significant differences among the visualized clusters.

Conclusion: Our findings suggest that the combination of the t-SNE visualization and multiple mean comparison through the PCA-projected exact F-test is a valuable tool for gene expression analysis. It not only enhances the interpretability of high-dimensional data but also provides a rigorous statistical framework for validating the observed patterns.

Keywords: Dimension reduction, *F* -test, Gene expression data, Multiple mean comparison, t-SNE plot.

INTRODUCTION

Gene expression data measures the activity of thousands of genes simultaneously, providing insights into the functional elements of the genome and the under- lying biological processes. These data sets, often generated by technologies such as microarrays or RNA sequencing (RNA-seq), have revolutionized the fields of genomics and molecular biology. By examining gene expression patterns, researchers can identify differentially expressed genes, understand cellular responses to various conditions, and uncover molecular mechanisms of diseases. The analysis of gene expression data has led to significant breakthroughs in understanding the molecular basis of diseases. For example, cancer research has greatly benefited from gene expression profiling. By comparing the gene expression patterns of cancerous and normal tissues, researchers can identify genes that are upregulated or downregulated in tumors, shedding light on the molecular mechanisms driving cancer progression. These differentially expressed genes can serve as potential biomarkers for diagnosis, prognosis, and therapeutic targets.

The high dimensionality of gene expression data poses significant challenges for analysis and interpretation. Each sample is characterized by the expression levels of thousands of genes, leading to complex data structures that are difficult to visualize

*Address correspondence to this author at the Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, 2000 Jintong Road, Tang Jia Wan, Zhuhai 519087, China; E-mail: jiajuanliang@uic.edu.cn

E-ISSN: 1929-6029/25

and analyze using traditional methods. Dimensionality reduction techniques are essential tools that help simplify these complex datasets while preserving their most informative aspects [1-2]. One commonly used dimensionality reduction method is Principal Component Analysis (PCA [3]). But it is well known that PCA is more suitable for linear dimension reduction. It is not desirable for nonlinear dimension reduction as is the case for high-dimensional gene expression data, which usually displays non-linear structure of manifold. Another popular method is the t-distributed stochastic neighbor embedding (t-SNE [4]), which is particularly effective for visualizing high-dimensional data in two or three dimensions [5].

The t-SNE minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects in the high-dimensional space and a distribution that measures pairwise similarities of the corresponding low-dimensional points. The t-SNE aims to map high-dimensional data points into a lower-dimensional space (usually two or three dimensions) while preserving the local structure of the data. This technique is well-suited for identifying complex structures in gene expression data, such as clusters or outliers, and has been widely used in single-cell RNA sequencing studies to reveal the heterogeneity of cell populations [6].

When applied to gene expression data, the t-SNE can reveal natural groupings or clusters of samples based on their gene expression profiles. These clusters often correspond to distinct biological states or cell types, making the t-SNE a valuable tool for exploratory data analysis in genomics. For instance, t-SNE plots have been used to identify and visualize different subpopulations of cells in single-cell RNA-seq data [7], aiding in the discovery of novel cell types and states. Once the t-SNE has been used to reduce the dimensionality of the gene expression data and visualize clusters, the next step is often to analyze these clusters further. One common approach is to perform clustering on the t-SNE reduced data to formally define groups of samples. Methods such as *k*-means clustering, hierarchical clustering, or density-based clustering can be employed to partition the data into distinct clusters.

After identifying clusters, researchers are typically interested in comparing the means of gene expression levels across these clusters to identify differentially expressed genes. This process involves statistical tests to determine whether the mean expression levels of specific genes differ significantly between clusters. The identification of differentially expressed genes can provide insights into the underlying mechanisms driving the observed clustering. For instance, certain genes might be upregulated in one cluster but downregulated in another, indicating their role in specific biological processes or disease states. By comparing mean expression levels, researchers can prioritize genes for further investigation, such as functional validation studies or therapeutic target identification [8-9].

To effectively compare gene expression levels across clusters, a variety of statistical methods can be employed. Commonly used techniques include Student's *t*-tests, ANOVA (analysis of variance), and more sophisticated methods like generalized linear models and mixed-effects models. These methods account for variability within and between clusters, ensuring that the detected differences are statistically robust and biologically meaningful. However, the application of these methods requires careful consideration of multiple testing issues, as the large number of genes analyzed simultaneously can lead to increased false discovery rates. Adjustments such as the Bonferroni correction or the False Discovery Rate (FDR) control are often applied to mitigate this problem [10-12].

Visualization of clusters and their corresponding gene expression profiles can greatly aid in the interpretation of the results. t-SNE plots are a popular tool for visualizing high-dimensional gene expression data in a two-dimensional space. By projecting complex data into a more interpretable format, t-SNE plots facilitate the identification of clusters and the exploration of their characteristics. When combined with mean comparison techniques, t-SNE plots can highlight which genes contribute most to the differences between clusters, providing a powerful approach for data-driven discovery.

In conclusion, comparing the means of gene expression levels across clusters is a fundamental step in the analysis of gene expression data. This process not only identifies differentially expressed genes but also enhances our understanding of the biological significance of the clusters. By leveraging statistical methods and visualization tools like the t-SNE plots, researchers can gain deeper insights into the molecular underpinnings of the data, paving the way for new discoveries in genomics and beyond.

This paper is organized as follows. Section 2 gives a simple illustration for the t-SNE plot by using two real high-dimensional datasets from publicly accessible sources. In section 3, we propose the projection-type exact F -test for multiple mean comparison for the case of high dimension with a possible small sample size. Section 4 gives a simple Monte Carlo study on the performance of the projection-type exact F-test. Section 5 illustrates the application of the projection-type F -test for comparing clusters of gene expression data from their t-SNE plots. Some concluding remarks are given in the last section.

2. AN ILLUSTRATION OF THE T-SNE PLOT

Because the t-SNE plot depends on a careful choice of its parameters such as the perplexity, which typically ranges between 5 and 50, smaller values of perplexity emphasize local data structure, while larger values focus more on global data structure, implement a t-SNE plot for a high-dimensional gene expression dataset needs some practical considerations as follows.

1) Data preprocessing: normalize or standardize data, as the t-SNE is sensitive to the scale of input features; and remove or handle outliers to prevent them from dominating the visualization.
2) Dimensionality reduction: consider reducing dimensions with techniques like PCA before applying the t-SNE, especially for very high-dimensional data.
3) Interpretation and validation: use multiple runs to ensure the stability of the results; visualize the results with other techniques (e.g., clustering) to validate the insights gained.
4) Computation resources: the t-SNE can be computationally intensive, especially for large datasets; utilize efficient implementations and appropriate hardware resources like the “Rtsne” package.

By carefully tuning these parameters and considering the practical aspects, one can effectively use the t-SNE to visualize high-dimensional data in a lower-dimensional space.

After applying the t-SNE plot, the interpretation of the plot includes:

1) Cluster analysis: examine the plot to identify distinct clusters of samples. Clusters in the t-SNE plot suggest groups of samples with similar gene expression profiles;
2) Biological insight: relate the clusters to biological or experimental conditions. For example, clusters might correspond to different cell types, treatment conditions, or disease status;
3) Outliers: identify outliers or anomalies that may warrant further investigation. Outliers in the plot could indicate unique or rare samples with distinct gene expression patterns.

The t-SNE plots in the following two examples are direct implementations of the R-package “Rtsne” available from the R-website https://cran.r-project.org/ by running install.package(“Rtsne”) and then running library(Rtsne) under the R command window.

Example 1. The gene expression dataset consists of gene mapping data of 50 genes with 1097 gene expression levels. Dimension p = 50 and sample size n = 1097. The dataset has ID TCGA-BRCA.htseq fpkm-uq.tsv downloaded from https://ucsc-public-main-xena-hub.s3.us-east1.amazonaws.com/download/chin2006_public%2Fchin2006Exp_genomicMatrix.gz

We first employ the elbow method [13] to guess how many clusters are suitable for data clustering. The elbow plot below in Figure **1** suggests five clusters may be enough to classify the gene expression data, where each elbow indicates a sharp change in WCSS=within-cluster sum of squares. Although classification of more than five clusters still seems feasible, the t-SNE plots of more than five clusters will show too many overlapped observations across clusters as shown in Figure **3**.

The t-SNE plots for the dataset in example 1 with different choices of the hyper- parameter perplexity are given as follows. It is suggested that the hyper-parameter perplexity is usually taken between 5 and 50 [14]. For small datasets, perplexity can be taken between 5 and 40. There are 50 genes with 1097 observations for the dataset in example 1. Figure **2** shows that datasets can be classified into five relatively clear clusters for different choices of the perplexity parameter. Figure **3** shows that classification of six clusters for the same dataset under different choices of the perplexity parameter results in too many overlapped observations, which may be classified into more than one cluster. Classification of more than six clusters for the same dataset under different choices of the perplexity parameter also results in too many overlapped observations, which may be classified into

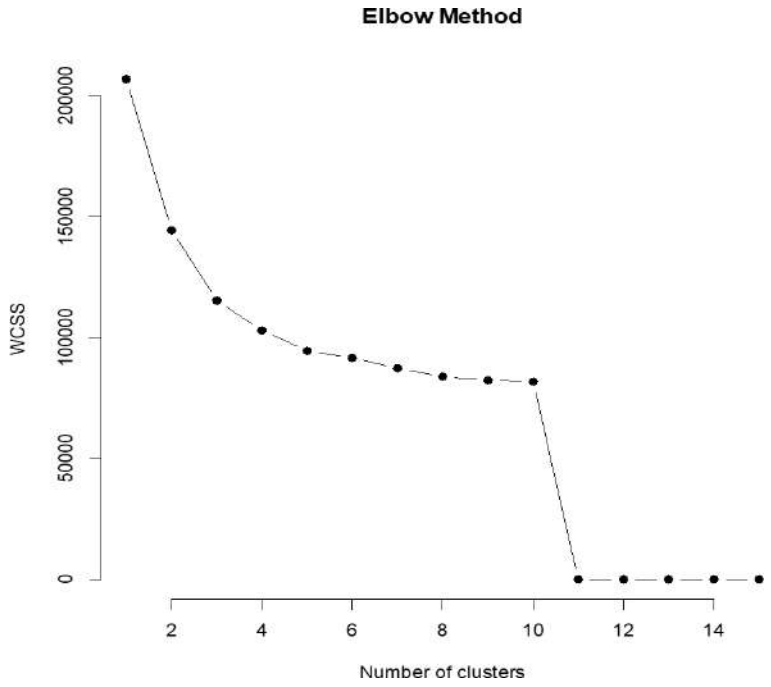

Figure 1: Elbow plot for example 1 dataset.

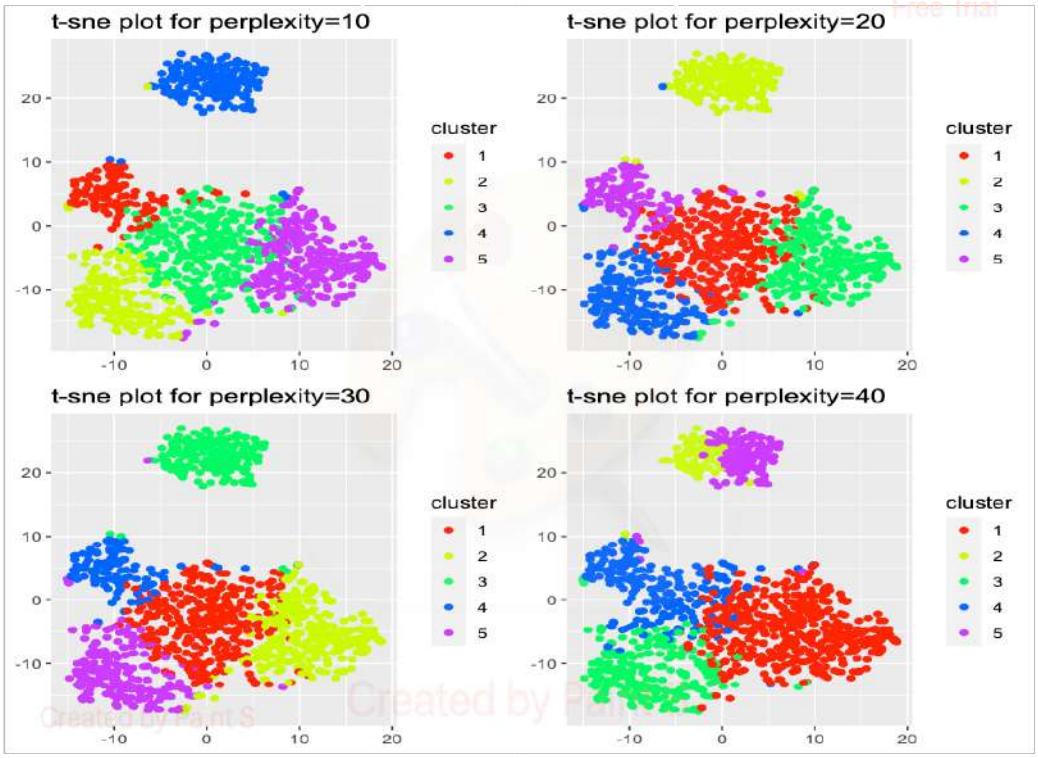

Figure 2: t-SNE plots with five clusters for example 1 dataset.

more than one cluster. Therefore, we choose classification of five clusters as a better option.

Example 2. The gene expression dataset consists of breast cancer gene expression data of 1217 genes with more than 65,000 gene expression levels. We only choose 1931 gene expression levels for illustration purpose. Dimension p = 1217 and sample size n = 1931. The dataset has ID TCGA-BRCA.htseq fpkm-uq.tsv downloaded from https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-BRCA.htseq_ fpkm-uq.tsv.gz

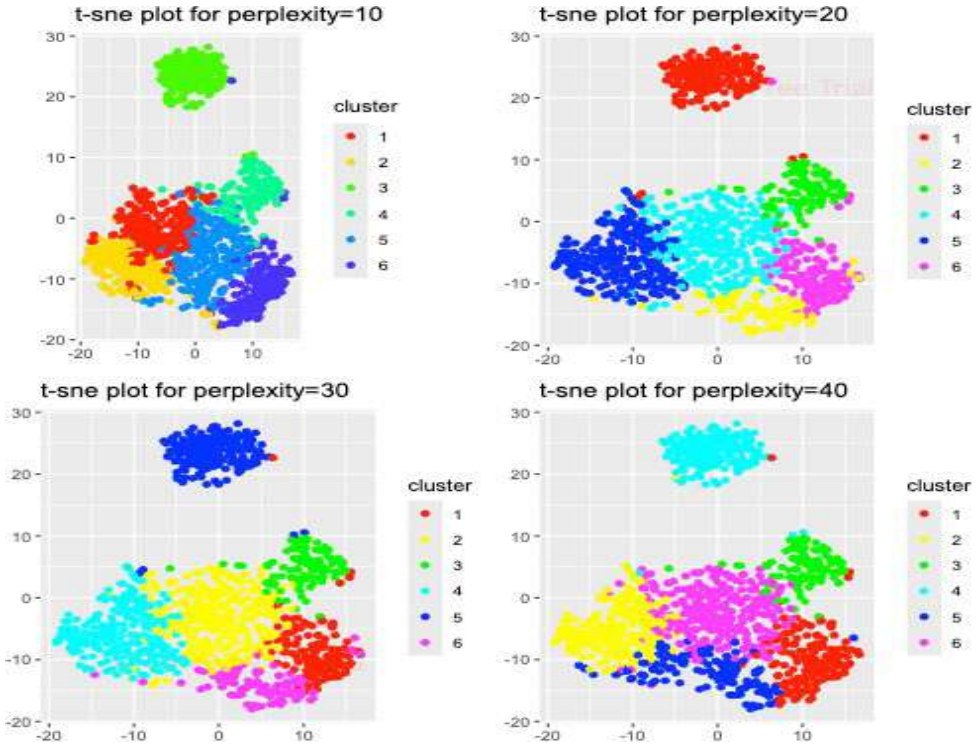

Figure 3: t-SNE plots with six clusters for example 1 dataset.

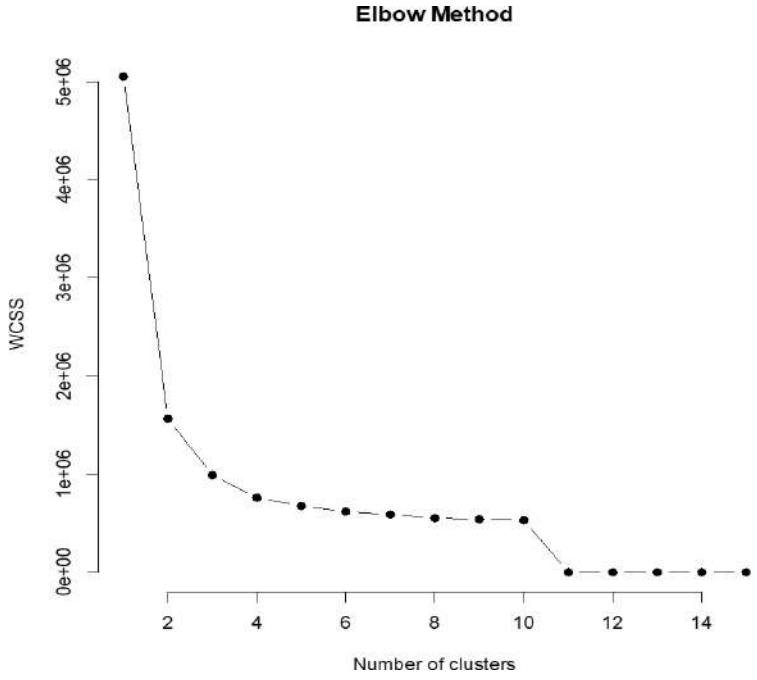

Figure 4: Elbow plot for example 2 dataset.

We first employ the elbow method to guess how many clusters are suitable for data clustering. The elbow plot below suggests three to six clusters may be suitable to classify the gene expression data, where each elbow indicates a sharp change in WCSS=within-cluster sum of squares. The best classification for the number of clusters may be re-evaluated by its associated t-SNE plots as follows.

For median datasets, perplexity can be taken between 20 and 50. There are 1217 genes with 1931 observations for the dataset in example 2. The t-SNE

plots for the dataset in example 2 with different choices of the hyper-parameter perplexity and different number of clusters are given in Figures **5-8**. It seems that a better option is to classify the data in to three clusters as shown in Figure 5. In Figure **6** with four clusters, the cluster with the smallest size can be combined into the cluster that is next to it. This results in three clusters. In Figure **7** with five clusters, the three clusters with the smaller sizes can be combined into one cluster. This also results in three clusters. In Figure **8** with six clusters, the four clusters with the smaller sizes can be combined into one cluster. This still results in three clusters. Based on the elbow plot in Figure **4** and re-evaluated by Figures **5-8**, it is feasible to classify the data in example 2 into three major clusters.

3. MULTIPLE MEAN COMPARISON THROUGH PCA DIMENSION REDUCTION

The t-SNE plot provides a visual observation on the clustering structure high- dimensional gene expression data. A quantitative comparison among the clusters from the t-SNE plot may help further understanding in the complex biological processes and diseases. With the advent of high-throughput sequencing technologies, researchers can now measure the expression levels of thousands of genes simultaneously. However, the sheer volume and complexity of this data necessitate sophisticated analytical techniques to extract meaningful insights. One crucial approach is the multiple mean comparison among clusters of gene expression data, which aids in identifying significant differences and similarities across different gene clusters.

The necessity of multiple mean comparison for analysis of gene expression data can be summarized in the following points:

1) Understanding biological variability: gene expression data often exhibit high variability due to biological differences, technical noise, and experimental conditions. By clustering genes with similar expression patterns, researchers can reduce this complexity. Multiple mean comparison among these clusters helps in discerning whether the observed differences in gene expression are statistically significant or merely due to random variation.

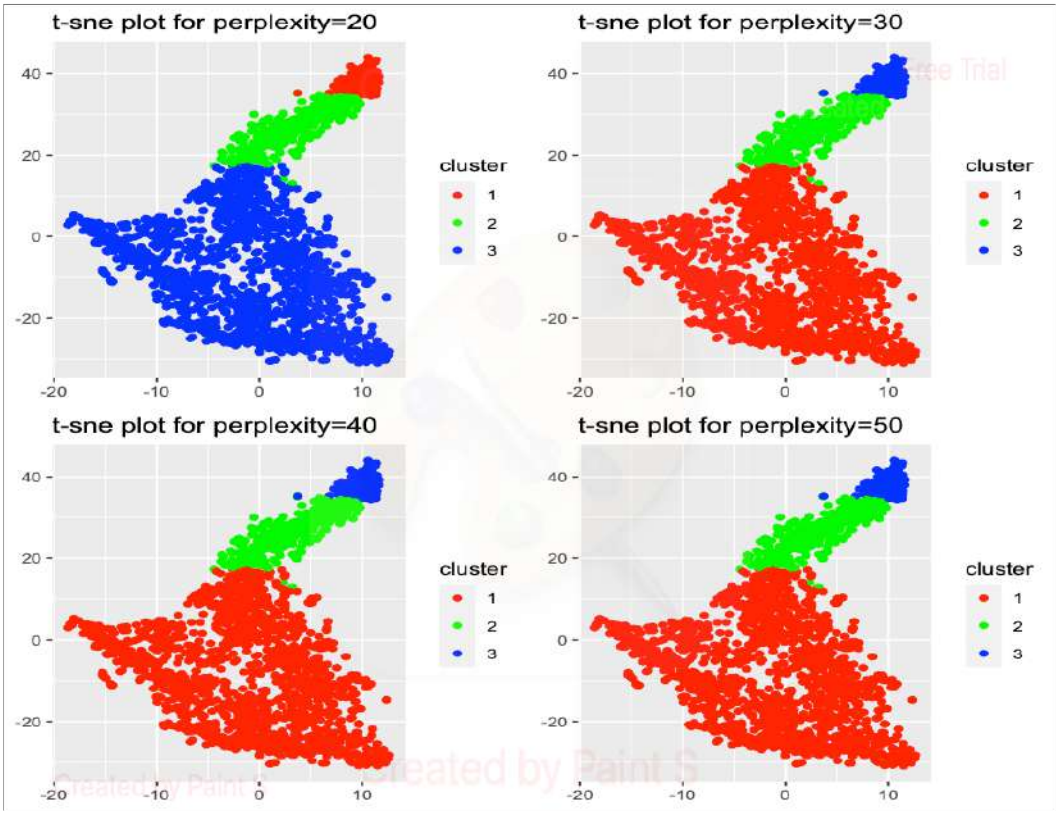

Figure 5: t-SNE plot with three clusters for example 2 dataset.

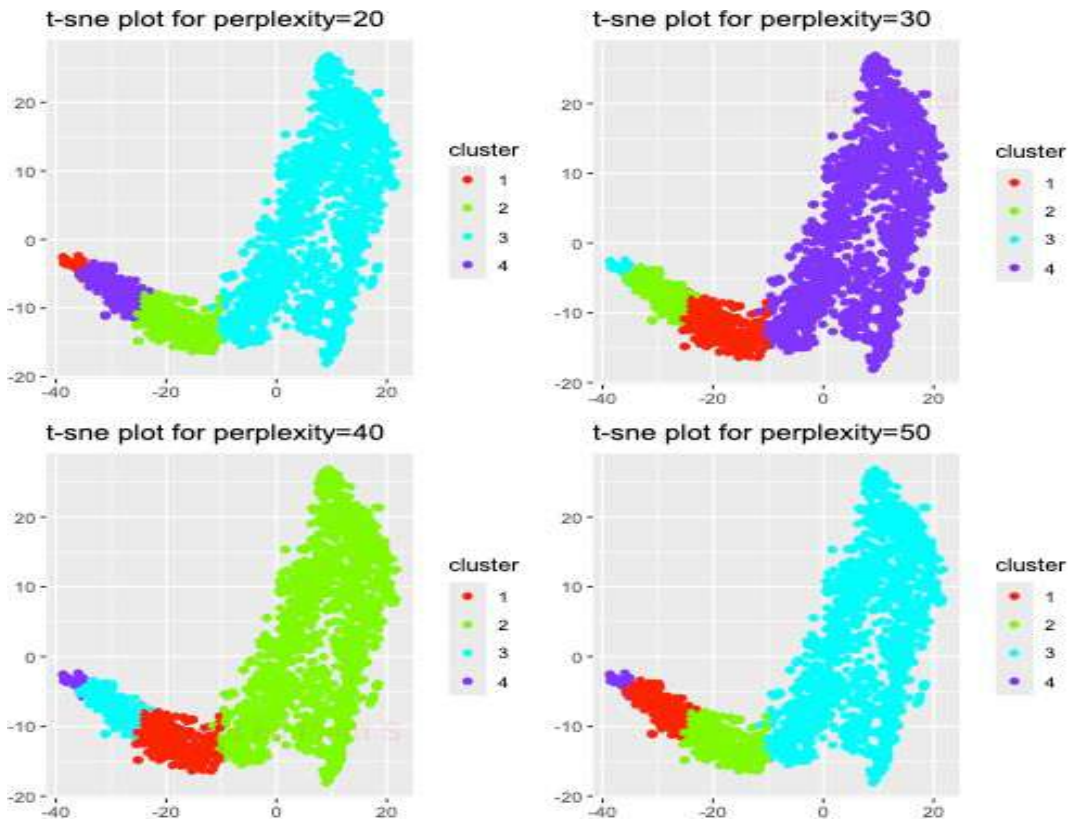

Figure 6: t-SNE plot with four clusters for example 2 dataset.

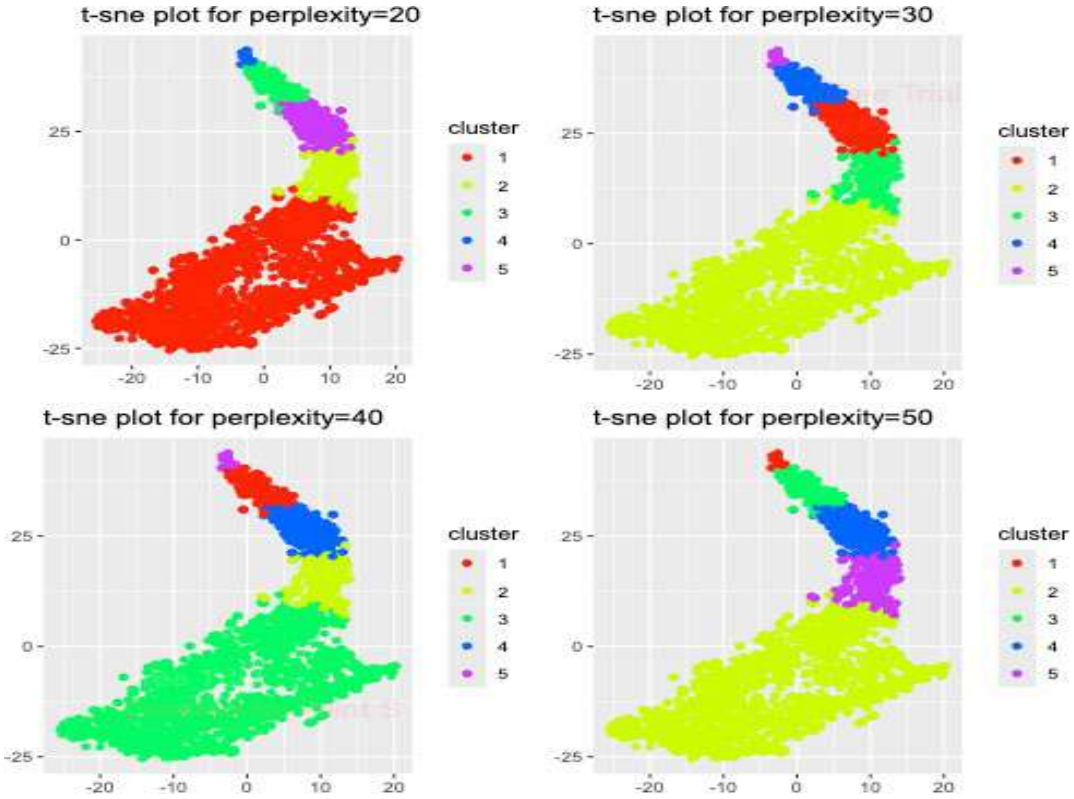

Figure 7: t-SNE plot with five clusters for example 2 dataset.

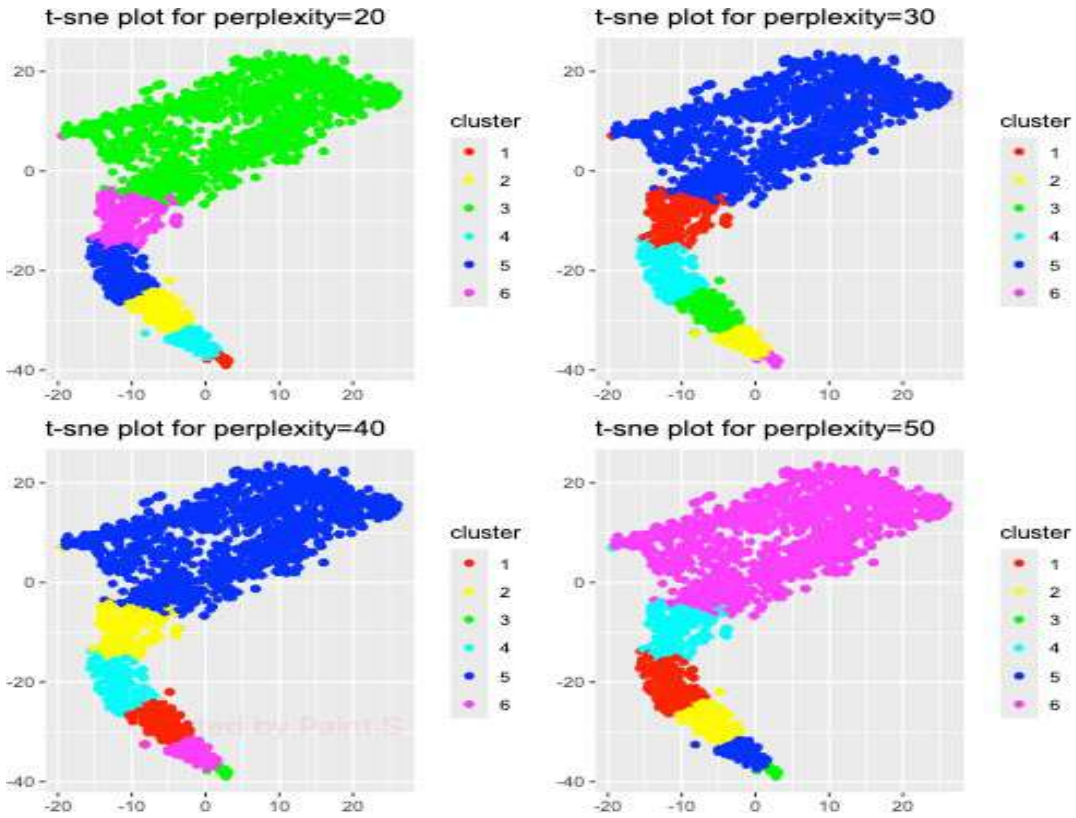

Figure 8: t-SNE plot with six clusters for example 2 dataset.

2) Identifying differentially expressed genes: in many studies, the goal is to identify genes that are differentially expressed under different conditions (e.g., diseased vs. healthy states). By comparing the means of gene expression levels across clusters, researchers can pinpoint specific genes or groups of genes that show significant changes, which might be indicative of underlying biological processes or disease mechanisms.

3) Enhancing data interpretation: high-dimensional data are often difficult to interpret directly. Clustering simplifies the data into more manageable subsets. Multiple mean comparisons among these clusters provide a clearer picture of the overall gene expression landscape, highlighting key differences and trends that can guide further biological investigation.

Gene expression studies typically involve testing thousands of genes simultaneously, leading to the problem of multiple mean comparisons. Once clusters are established, statistical methods such as ANOVA or Student's *t*-tests are used to compare the mean expression levels between clusters. Multivariate analysis of variance (MANOVA) is an extension to univariate ANOVA, it requires the total sample size (the number of expression levels) must be greater than the dimension (the number of genes). This limits the multiple mean comparison based on MANOVA. Some nonparametric methods like the permutation test [15] may be applied to multiple mean comparison. But nonparametric methods usually result in statistics whose finite-sample null distributions are unknown or difficult to obtain their critical values

In this section, we employ the idea of PCA to construct the Läuter-type *F*-test [16] for multiple mean comparison across clusters of gene expression data which are determined by the t-SNE plot. Let

$$\boldsymbol{\mu}_l = \text{the mean vector of cluster } l,\ l = 1, \ldots, k \quad (1)$$

and

$$\mathbf{x}_{lj} = (x_{lj1}, \ldots, x_{ljp})^t : p \times 1,\ j = 1, \ldots, n_l;\ l = 1, \ldots, k \quad (2)$$

be the observations from cluster l with sample size n_l, where the superscript "t" stands for the transpose of a row vector or a matrix. We assume normal samples $\{\boldsymbol{x}_{lj} : j=1, \ldots, n_l\}$ are i.i.d. p-dimensional normal $N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma})$ with equal covariance matrices. We want to test the hypothesis

$H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k$ (3)

against the alternative that at least two means are not equal.

Theorem 1. Under the notations in (1)-(3), let

$$X_1 = \begin{pmatrix} x_{11}^t \\ \vdots \\ x_{1n_1}^t \end{pmatrix}, n_1 \times p, \ldots, X_k = \begin{pmatrix} x_{k1}^t \\ \vdots \\ x_{kn_k}^t \end{pmatrix}, n_k \times p,$$

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}, n \times p, n = \sum_{i=1}^k n_i. \quad (4)$$

Under the null hypothesis (3), we define the Helmert's transformation [17]:

$$Y = AX: \ (n-1) \times n \quad (5)$$

with the constant matrix $\boldsymbol{A}$ defined by

$$A = (a_{ij}): (n-1) \times n, \ a_{ij} = \begin{cases} \frac{1}{\sqrt{i(i+1)}}, & j = 1, \ldots, i \\ \frac{-i}{\sqrt{i(i+1)}}, & j = i+1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Define eigenvalue-eigenvector problem

$$\frac{1}{n-1}(Y'Y) = D\Lambda \quad (7)$$

where $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_q)$, $p \times q$, $q = \min(n-1, p) - 1$. $\boldsymbol{D}$ consists of q eigenvectors $\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_q\}$ associated with q positive eigenvalues of the non-negative definite matrix $\boldsymbol{Y}^t\boldsymbol{Y}$. $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_q)$ consists of the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_q > 0$. Let

$\mathbf{D}_r = (\mathbf{d}_1, \ldots, \mathbf{d}_r) : p \times r, r = 1, \ldots, q.$ $\mathbf{Z}_r = \mathbf{Y}\,\mathbf{D}_r,$

$$H = Z_r^t \left(\frac{1}{n-1}\mathbf{1}_{n-1}\mathbf{1}_{n-1}^t\right) Z_r,$$
$$G = Z_r^t \left(I_{n-1} - \frac{1}{n-1}\mathbf{1}_{n-1}\mathbf{1}_{n-1}^t\right) Z_r, \quad (8)$$

where $\mathbf{1}_{n-1}$ stands for the vector of ones with dimension $(n-1) \times 1$, $\boldsymbol{I}_{n-1}$ for the identity matrix with dimension $(n-1) \times (n-1)$, $r = 1, \ldots, \min(n-1, q) - 1$. Define the statistic

$$F_r = \frac{n-1-r}{r} \text{trace}(HG^{-1}) = \frac{n-1-r}{(n-1)r} \mathbf{1}_{n-1}^t Z_r G^{-1} Z_r^t \mathbf{1}_{n-1} \quad (9)$$

Then under the null hypothesis (3), F_r has an F -distribution $F(r, n-1-r)$. **Proof**. Under the assumption of equal covariance matrices and the null hypothesis (3), the row vectors in the matrix $\boldsymbol{X}$ in (4) are i.i.d. with a p-dimensional normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(\boldsymbol{\mu} = \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k)$. The Helmert's transformation $\boldsymbol{Y} = \boldsymbol{AX}$ in (5) with $\boldsymbol{A}$ given by (6) being a row-orthogonal matrix $\boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{I}_{n-1}$ and $\boldsymbol{A}\mathbf{1}_n = \mathbf{0}$. Using the notation of matrix normal distribution and the Kronecker product "$\otimes$" and "$=^d$" meaning the two sides of the equality have the same probability distribution, we can write the distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$X =^d vec(X^t) \sim N_{n \times p}(1_n \otimes \mu, I_n \otimes \Sigma).$$

$$Y =^d vec(Y^t) = vec[(AX)^t] = (A \otimes I_p) vec(X^t) \sim N_{n \times p}(1_n \otimes \mu, I_n \otimes \Sigma).$$

$$E(vecY^t) = (A \otimes I_p) E[vec(X^t)] = (A \otimes I_p)(1_n \otimes \mu) = (A1_n) \otimes (I_p \mu) = 0 \otimes \mu = 0.$$

$$cov[(vecY^t)] = (A \otimes I_p) cov[vec(X^t)] (A \otimes I_p)^t$$

$$= (A \otimes I_p)(I_n \otimes \Sigma)(A^t \otimes I_p) = (AA^t) \otimes \Sigma = I_{n-1} \otimes \Sigma.$$

This means that the row vectors in $\boldsymbol{Y}$ are i.i.d. with a normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{Y}$ satisfies the conditions in Theorem 1 of Läuter (1996) [16]. The random matrix $\boldsymbol{Z}_r$ in (8) has a left-spherical matrix distribution with $P(\boldsymbol{Z}_r = \mathbf{0}) = 0$ [18]. According to Theorem 1 of Läuter (1996), the statistic F_r in (9) has an F -distribution $F(r, n-1-r)$. This completes the proof.

Theorem 1 provides an F -test for hypothesis (3). Reject hypothesis (3) at a given level $0 < \alpha < 1$ if $F_r > F(1-\alpha; r, n-1-r)$, here $F(1-\alpha; r, n-1-r)$ stands for the $100(1-\alpha)\%$-percentile of the F -distribution $F(r, n-1-r)$. Here $r = 1, \ldots, \min(n-1, q) - 1$ with q=the number positive eigenvalues in the matrix $\boldsymbol{Y}^t\boldsymbol{Y}$ $(p \times p)$. r can be considered as the projection dimension when projecting the Helmert-transformed data in $\boldsymbol{Y}$ onto the PCA directions determined by $\boldsymbol{Z}_r$ in (8). A test for (3) based on the F_r in (9) is called the projected F -test, which is an exact F -test under the null hypothesis (3).

4. A SIMPLE MONTE CARLO STUDY

The performance of the F_r-statistic (9) can be partially viewed by a simple Monte Carlo study through choosing different projection dimensions and comparing the F_r-test with the classical Wilks' Λ-test in MANOVA. The Λ-statistic is defined by

$$\Lambda = \frac{|W|}{|W+B|} \sim \Lambda(p, n-k, k-1) \quad (10)$$

under the null hypothesis (3), where

$$W = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^t \text{ and } B = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^t \quad (11)$$

are respectively the "within-samples" and "between-samples". *W* characterizes the total variation from the samples, *B* characterizes the total variation from the differences between the populations (treatments), and

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=}^{n_i} x_{ij}\,,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\,,\ \ n = \sum_{i=1}^{k} n_i. \quad (12)$$

Here $\bar{x}_i$ stands for the i-th sample mean for the sample from the i-th treatment and $\bar{x}$ for the overall sample mean from all samples. The distribution $\Lambda(p, n-k, k-1)$ in (10) is called the Wilks' Λ-distribution [19]. Hypothesis (3) is rejected for small values of Λ.

The exact distribution of the Wilks' Λ-statistic (10) is available only for the special cases of k =2 and k =3 (two or three treatments, or two or three populations [19] (page 83):

$$k=2:\ \frac{n-p-1}{p}\cdot\frac{1-\Lambda(p,n-2,1)}{\Lambda(p,n-2,1)} \sim F(p,n-p-1),$$

$$k=3:\ \frac{n-p-2}{p}\cdot\frac{1-\Lambda^{\frac{1}{2}}(p,n-3,2)}{\Lambda^{\frac{1}{2}}(p,n-3,2)} \sim F\big(2p,2(n-p-2)\big). \quad (13)$$

For the general case of k, the asymptotic χ^2-distribution is employed when using the Wilks' Λ-statistic (10):

$$-\Big[n-k-\frac{1}{2}(p-k+2)\Big]\log\Lambda(p,n-k,k-1) \rightarrow \chi^2\big((k-1)p\big),\ n\rightarrow\infty \quad (14)$$

under the null hypothesis (3). Therefore, the classical method for MANOVA based on the Wilks' Λ-statistic (10) is mainly for the case of large total sample size n.

Because the null distributions of the F_r-statistic (9) and the Wilks'Λ-statistic (10) do not depend on the normal means and the covariance matrix under the null hypothesis (3), we choose the following sample designs for the mean vectors and covariance matrices from the populations for multiple mean comparison in our Monte Carlo study.

Design 1: k =2 groups and dimension p = 10, 20, 30 with sample sizes $n_1 = n_2 = 20$

with

$$\mu_1 = 1_p,\ \mu_2 = d1_p,\ \Sigma = \left(\rho_{ij}\right),\ p\times p, \quad (15)$$

where the constant d controls the difference between the two mean vectors, and

$$\rho_{ij} = \begin{cases} 1, & i=j \\ 0.5, & i\neq j \end{cases} \quad (16)$$

for $i,j = 1,\ldots,p$.

Design 2: k =3 groups and dimension p = 10, 20, 30 with sample sizes $n_1 = n_2 = n_3 = 20$ with

$$\boldsymbol{\mu}_1 = \mathbf{1}_p,\ \boldsymbol{\mu}_2 = d\mathbf{1}_p,\ \boldsymbol{\mu}_3 = 2d\mathbf{1}_p \quad (17)$$

with the same covariance ρ_{ij} as given in (16).

Design 3: k =5 groups and dimension p =10, 20, 30 with sample sizes $n_1 = \ldots = n_5$ =20 with

$$\boldsymbol{\mu}_i = (i-1)d\mathbf{1}_p,\ i = 1,\ldots,10. \quad (18)$$

and the same covariance ρ_{ij} as given in (16).

Design 4: k =10 groups and dimension p =10, 20, 30 with sample sizes $n_1 = \ldots = n_{10} = 20$ with the same mean and covariance structure as in Design 3.

The multivariate normal samples are generated from each of the above designs for 2,000 replications. The empirical power for each design is computed by counting the relative frequency using the significance level α =0.05. The null distribution for the Wilks-test is an exact F -distribution for k =2 and k =3 as given by (13), it is an asymptotic chi-square distribution for k =5 and k =10 as given by (14). Large values of the statistics in (13) or (14) imply rejection of hypothesis (3). The power comparison for the above four designs is illustrated by Figures **9-10**, where the exact F -tests are from the PCA projected F -test given by (9) with different projection dimensions r_1=[r/4], r_2 =[r/3], r_3 =[r/2], and r_4 =[3r/4], here [·] stands for the integer part of a real number. Figures **9-10** show that the classical Wilks' Λ-test for multiple mean comparison may perform equally well or even better than the PCA-projection tests in low dimensional cases with small number of groups. It is obvious that Wilks' Λ-test begins losing power with the increase of data dimension or the number of groups. The PCA-projection tests seem to perform better for projection dimension between r_1 =[r/4] and r_2=[r/3] with r=min(n−1,p)−1, here n stands for the total sample size from all groups and p=data dimension.

5. ILLUSTRATIVE EXAMPLES

Example 3. (Example 1 continued) The clusters are displayed by the t-SNE plot in Figure **2**. We carry out the projected F-test for multiple mean comparison among the five clusters in Figure **2**. This is to test the hypothesis (3) with k=5. we carry out the projection F-test F_r in (9) by choosing four projection dimensions as

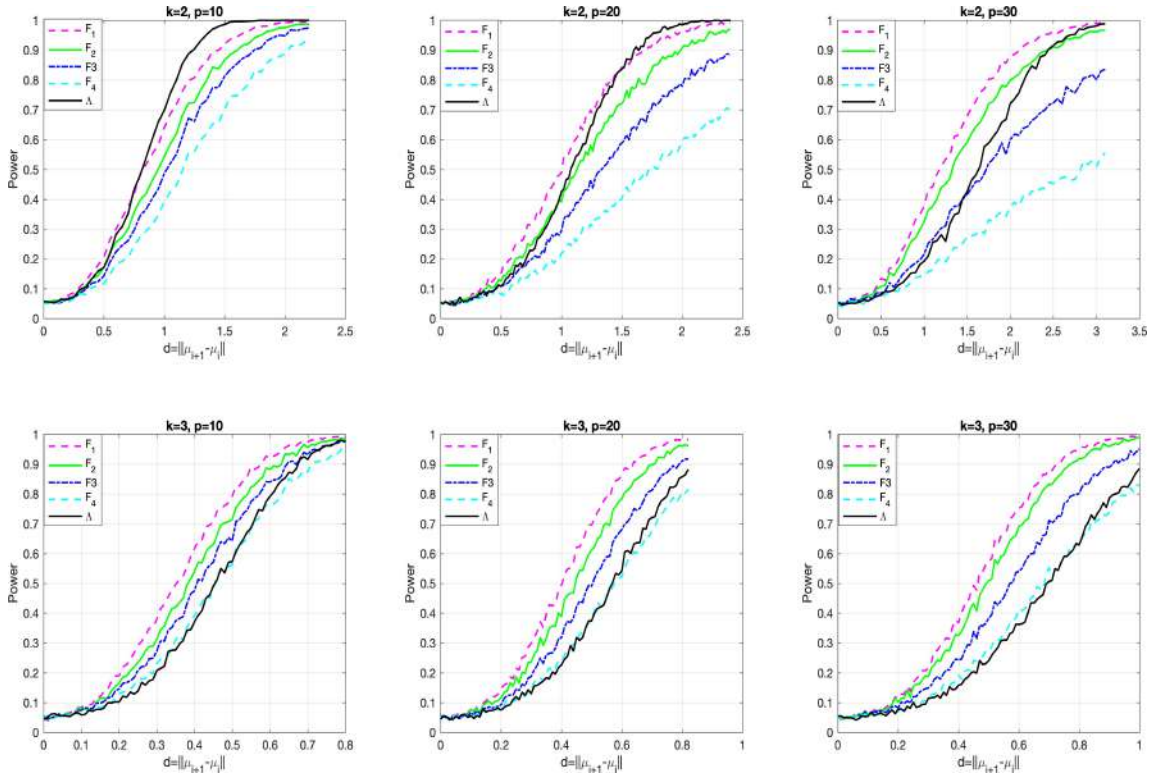

Figure 9: Power comparison among five tests: $F_1=F_{r1}$, $F_2=F_{r2}$, $F_3=F_{r3}$, $F_4=F_{r4}$, and Λ=Wilks' Λ.

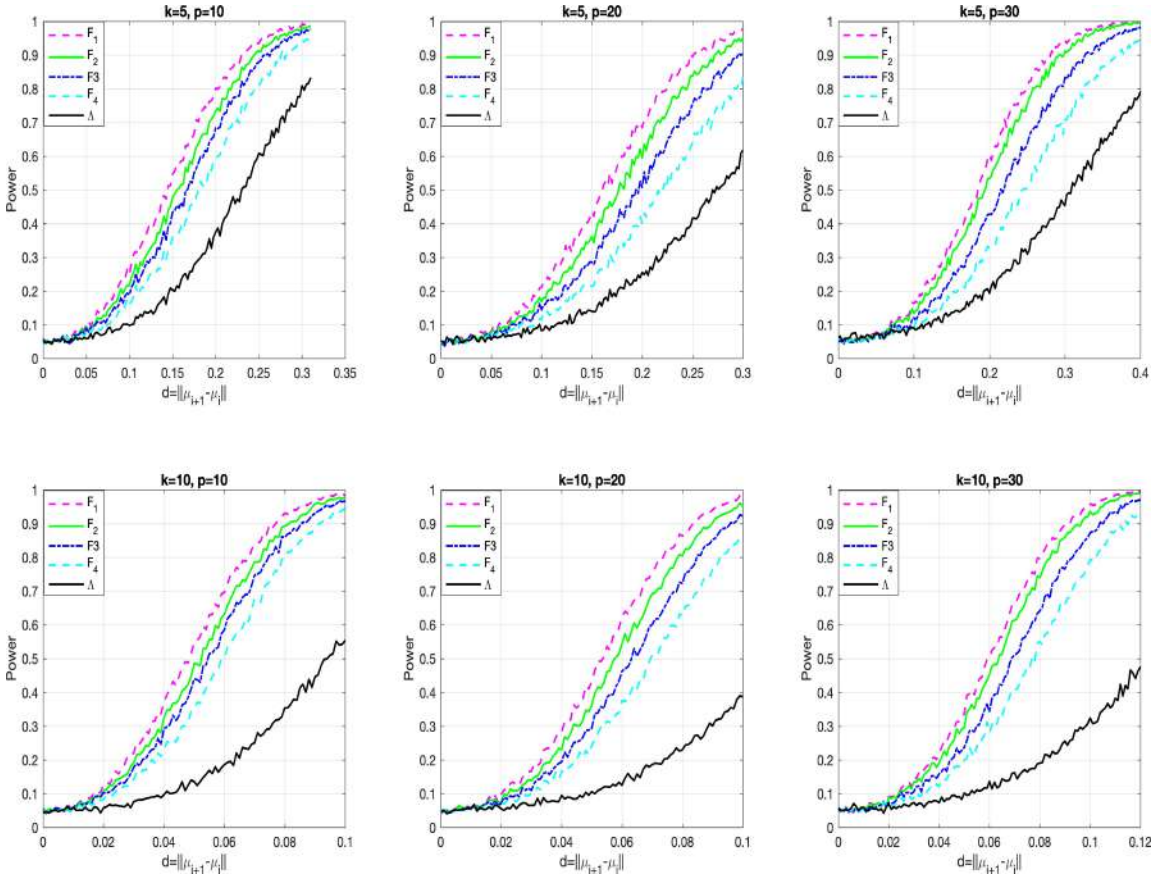

Figure 10: Power comparison among five tests: $F_1=F_{r1}$, $F_2=F_{r2}$, $F_3=F_{r3}$, $F_4=F_{r4}$, and Λ=Wilks' Λ.

Table 1: Projected *F*-tests for the dataset in Example 1

Projection dimension r	r_1=12	r_2=16	r_3=24	r_4=36
Projected *F*-distribution	F (12, 1084)	F (16, 1080)	F (36, 1072)	F(36,1060)
Projected *F*-value	97.1724	78.9793	54.7047	37.9098
p-value	1.9e-162	7.6e-169	1.6e-163	1.62e-163
	Wilks' approx. $\chi^2(200)$=101068.4, *p*-value=0.0000			

Note: In Table **1**, numbers like 1.9e-162 means = 1.9×10^{-162}.

the same as in Figures **9-10** and the Wilks' Λ-test (14). The test results are summarized in Table **1** below. The *p*-values show that there exists highly significant differences among the mean expression levels across the five clusters in Figure **2**.

Example 4. (Example 2 continued) The clusters are displayed by the t-SNE plot in Figure **4**. We carry out the projected *F* -test for multiple mean comparison among the three clusters in Figure **4**. This is to test the hypothesis (3) with k=3. We choose the first few PCA directions that can explain at least 80% of the variation in the PCA equation (8). It turns out that the first PCA direction already contributes more than 99.96% of the variation. Therefore, we only choose the first PCA direction for the projected *F*-test, which is $F(1,1929)$, it has a *p*-value=2.38888e-63≈0. It shows that there exists highly significant differences among the mean expression levels across the three clusters in Figure **4**. In this example, the data dimension (the number of genes) p=1217 is relatively close to the total sample size n=1931. The Wilks' Λ-test can be still applied but may be much less powerful than the one-dimensional PCA projected $F(1,1929)$-test. Actually, Λ=0 in this example, which implies no within-group variation. So the Wilks' Λ-test doesn't make much sense.

6. CONCLUDING REMARKS

In this paper, we presented a methodology for multiple mean comparison of clusters derived from gene expression data using the t-SNE plots. The application of t-SNE in visualizing high-dimensional gene expression data has proven to be a powerful tool for uncovering inherent structures and patterns that traditional methods might overlook. By leveraging t-SNE, we effectively reduce the dimensionality of the data while preserving the local and global data structure, facilitating more intuitive and informative visualizations.

Through our analysis, we aimed to address the challenge of interpreting gene ex- pression clusters by applying multiple mean comparison techniques. This approach allows us to statistically validate the differences between clusters, providing a robust framework for identifying biologically significant patterns. The integration of the t-SNE visualization with statistical testing bridges the gap between exploratory data analysis and confirmatory analysis, ensuring that the observed patterns are not merely artifacts of the visualization process.

The two real gene expression data examples demonstrated the practical application of our methodology. In both real-data examples, the t-SNE plots revealed distinct clusters corresponding to different expression levels. Multiple mean comparison tests further confirmed the significant differences between these clusters. The real-data examples 3-4 show that it is very common that the sample size may be less than the data dimension. This makes the classical Wilks' Λ-test inapplicable for testing high-dimensional mean under the multivariate normal assumption. Other available nonparametric tests in the literature for this purpose require large- sample theory to obtain the null distributions of the test statistics. Therefore, our proposed projected *F*-test shows some superiority to some existing approaches to high-dimensional multiple mean comparison.

Our findings suggest that the combination of the t-SNE visualization and multiple mean comparison is a valuable tool for gene expression analysis. It not only enhances the interpretability of high-dimensional data but also provides a rigorous statistical framework for validating the observed patterns. This approach can be extended to various types of genomics data, offering a versatile solution for complex biological data analysis.

However, it is essential to acknowledge certain limitations of our methodology. The performance of t-

SNE can be sensitive to the choice of parameters, such as perplexity and learning rate. Careful parameter tuning is necessary to achieve meaningful visualizations. Additionally, the multiple mean comparison tests assume equal covariance matrices across all clusters, which may not be the case in some applications. The multiple mean comparison tests also assume that the data within each cluster follows a normal distribution, which may not always be the case. Future work could explore robust statistical methods that relax these assumptions and improve the reliability of the results.

It should be pointed out that t-SNE plot for clustering high-dimensional gene expression data and PCA for dimension reduction are established methods in the literature. Our contribution in this paper is to double validate the t-SNE plot by the generalized F-test derived from PCA, which makes it possible to compare the mean difference for the situation of high dimension with a possible small total sample size. This kind of multiple mean comparison cannot be tested by the traditional MANOVA approach. There are established methods for identifying differentially expressed genes from single-cell gene expression datasets, see, for example, references [20-22]. Gene expression data can be also analyzed by examining differential expression of replicated count data, and some R packages are available [23]. It will be a big project to compare different methods for analysis of different types of gene expression data.

In conclusion, our study highlights the effectiveness of combining t-SNE plots with multiple mean comparison for analyzing gene expression data. This integrated approach may facilitate the discovery of biologically meaningful patterns and pro- vides a solid statistical foundation for validating these findings. We anticipate that our methodology will be beneficial for researchers in genomics and other fields where high-dimensional data analysis is crucial. Future research could focus on optimizing parameter selection for the t-SNE plots and exploring alternative statistical methods to further enhance the robustness and applicability of our approach.

ACKNOWLEDGEMENT

This work was supported in part by the Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College (UIC), project codes UICR0400026-21 and 2022B1212010006.

REFERENCES

[1] Roweis ST, Saul KL. Nonlinear dimensionality reduction by locally linear embedding. Science 2000; 290: 2323-2326. https://doi.org/10.1126/science.290.5500.2323

[2] Tenenbaum JB, Silva VD, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000; 290: 2319-2323. https://doi.org/10.1126/science.290.5500.2319

[3] Jolliffe IT. Principal Component Analysis. Springer, New York, 1986. https://doi.org/10.1007/978-1-4757-1904-8

[4] van der Maaten L,Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research 2008; 9: 2579-2605.

[5] Konstorum A, Jekel N, Vidal E, Laubenbacher R. Comparative analysis of linear and nonlinear dimension reduction techniques on mass cytometry data. BioRxi 2018. https://doi.org/10.1101/273862

[6] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Regev A. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015; 161(5): 1202-1214. https://doi.org/10.1016/j.cell.2015.05.002

[7] Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology 2013; 31(6): 545-552. https://doi.org/10.1038/nbt.2594

[8] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology 2015; 33(5): 495-502. https://doi.org/10.1038/nbt.3192

[9] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Regev A. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015; 161(5): 1202-1214. https://doi.org/10.1016/j.cell.2015.05.002

[10] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society (Ser. B) 1995; 57(1): 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[11] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statistical Science 2003; 18(1): 71-103. https://doi.org/10.1214/ss/1056397487

[12] Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19(3): 368-375. https://doi.org/10.1093/bioinformatics/btf877

[13] Ketchen DJ, Shook CL. The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal 1996; 17(6): 441-458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

[14] Wattenberg M, Vi´egas F, Johnson I. How to Use t-SNE Effectively. Distill 2016. https://doi.org/10.23915/distill.00002

[15] Good PI. Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer 2005.

[16] Läuter J. Exact *t* and *F* tests for analyzing studies with multiple end-points. Biometrics 1996; 52(3): 964-970. https://doi.org/10.2307/2533057

[17] Mardia KV. Tests of univariate and multivariate normality. Krishnaiah PR. ed. Handbook of Statistics, North-Holland Publishing Company 1980; 1: 279-320. https://doi.org/10.1016/S0169-7161(80)01011-5

[18] Fang KT, Zhang Y. Generalized Multivariate Analysis. Springer-Verlag and Science Press, Berlin/Beijing 1990.

[19] Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. Academic Press, London and New York 1979.

[20] Junttila S, Smolanda J, Elo, LL. Bench marking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. Briefings in Bioinformatics 2022; 23(5): 1-14. https://doi.org/10.1093/bib/bbac286

[21] Gezelius H, Enblad AP, Lundmark A, Aberg M, Blom K, Rudfeldt J, Raine A, Harila A, Rendo V, Heinäniemi M, Andersson C, Nordlund J. Comparison of high-throughput single-cell RNA-seq methods for ex vivo drug screening. NAR Genomics and Bioinformatics 2024; 6: 1-13. https://doi.org/10.1093/nargab/lqae001

[22] Gao X, Hu D, Gogo L, Li H. ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions. Bioinformatics 2019; 35(17): 3038-3045. https://doi.org/10.1093/bioinformatics/btz024

[23] Seyednasrollah F, Laiho A, Elo, LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics 2013; 16(1): 59-70. https://doi.org/10.1093/bib/bbt086

Received on 07-12-2024 Accepted on 02-01-2025 Published on 22-01-2025

https://doi.org/10.6000/1929-6029.2025.14.01