SciMedCentral

Annals of Biometrics & Biostatistics

Research Article

A t-SNE-Plot-Aided Generalized F-test for Analysis of Gene Expression Data

Jeffrey Liang[1]*, Hong-Bin Fang[2] and Yongsheng Bai[3]

[1]*Daniel Hand High School, Madison, CT 06443, USA*

[2]*Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057, USA*

[3]*Next-Gen Intelligent Science Training, Ann Arbor, MI 48105, USA*

***Corresponding author**

Jeffrey Liang, Danniel Hand High School, Madison, CT 06443, USA, 286 Green Hill Road, Madison, CT 06443, USA, Tel: +1 475-355-3112

Submitted: 07 October 2024

Accepted: 22 October 2024

Published: 24 October 2024

ISSN: 2333-7109

Keywords

- Dimension reduction; Gene expression data; Multiple mean comparison; tSNE plot; Generalized F-test

Abstract

In this paper, we introduce a novel approach for analyzing gene expression data by integrating the t-distributed Stochastic Neighbor Embedding (t-SNE) for data clustering with a generalized F-test for multiple mean comparison. High-dimensional gene expression data often poses challenges when the number of features exceeds the total sample size from individual clusters, limiting the applicability of traditional multivariate methods such as Multivariate Analysis of Variance (MANOVA). By employing t-SNE, we first perform nonlinear dimensionality reduction to cluster gene expression data, providing clear visual separation of different groups. Following this, a generalized F-test is applied to compare the mean expression levels across these clusters. The method is further enhanced through projections onto lower dimensions using Principal Component Analysis (PCA), ensuring robustness across different projection spaces. Our approach provides an efficient solution to the problem of multiple mean comparison in high-dimensional settings, where traditional methods fall short. We demonstrate the effectiveness of the proposed method through a case study involving real gene expression data, highlighting its practical utility for researchers in genomics and bioinformatics. Future work will explore post-hoc analyses after rejecting the null hypothesis of equal mean expression levels.

ABBREVIATIONS

IID: Independent Identically Distributed; MANOVA: Multivariate Analysis of Variance; PCA: Principal Component Analysis; t-SNE: t-Stochastic Neighbor Embedding

INTRODUCTION

The analysis of gene expression data has become increasingly complex with the advent of high-throughput technologies, which generate large-scale, high-dimensional data. One of the challenges associated with this type of data is the need to accurately classify and compare gene expressions across different experimental conditions or samples. Traditional multivariate statistical methods, such as Multivariate Analysis Of Variance (MANOVA), are often limited by the curse of dimensionality [1], particularly when the number of dimensions exceeds the number of observations. In such cases, standard methods may fail to deliver meaningful insights due to the overfitting of models, leading to unstable or unreliable results. To address these challenges, novel approaches that combine nonlinear dimensionality reduction with robust statistical tests are becoming essential [2,3]. This paper introduces a method that integrates the t-distributed Stochastic Neighbor Embedding (t-SNE, [4]) technique with the generalized F-test [5] for the analysis of gene expression data. The primary aim is to utilize the strength of t-SNE for classifying high dimensional gene expression data into meaningful clusters and subsequently applying the generalized F-test to compare means across these clusters. The method capitalizes on the nonlinear dimensionality reduction capability of t-SNE to visualize complex gene expression data structures, followed by the application of a statistical test suitable for situations where traditional methods such as MANOVA are not applicable due to the data's high dimensionality relative to sample sizes.

The t-SNE algorithm is particularly advantageous in high-dimensional data contexts as it effectively preserves the local structure of the data while allowing for a global arrangement that reflects meaningful clusters. This method is especially useful for visualizing gene expression data, where intricate relationships between genes can be difficult to interpret using standard dimensionality reduction techniques. By mapping highdimensional gene expression profiles into a lower-dimensional space, t-SNE helps reveal natural groupings of gene expressions, which can be crucial for identifying distinct biological processes or experimental conditions.

After visualizing and identifying clusters with t-SNE, the generalized F-test is employed to perform multiple mean comparisons across the clusters. The generalized F-test has been recognized as a robust alternative to traditional methods when dealing with highdimensional data, particularly in cases where the dimensionality exceeds the total number of samples. Unlike classical MANOVA, which is limited by the requirement that the total sample size must exceed the number of variables, the generalized F-test accommodates the high-dimensional setting, allowing for the comparison of multiple group means even when the data's dimensionality is greater than the total sample sizes from individual clusters. This approach offers several benefits for the analysis of gene expression data. First, the use of t-SNE allows for the visualization of complex gene expression patterns, providing a clear and interpretable view of the data's structure. This is particularly valuable for understanding the underlying biological processes and for identifying distinct groups of gene expressions that may correspond to different functional pathways or disease states [6,7]. Second, the generalized F-test enables robust statistical inference in high-dimensional settings, avoiding the limitations of traditional multivariate analysis techniques. By combining these two methods, researchers can gain both an intuitive visual understanding of their data and a rigorous statistical framework for comparing group means.

In summary, this paper presents a novel approach that combines the t-SNE plot's clustering and visualization capabilities with the generalized F-test's statistical rigor for analyzing high-dimensional gene expression data. The proposed method allows for effective classification of gene expressions and robust comparison of means across clusters, making it particularly suitable for applications in genomics where traditional methods may falter due to the high dimensionality of the data. Section 2 presents the basic idea of the t-SNE plot and its illustration through real datasets. Section 3 gives simple introduction to the generalized F-test and its implementation on real datasets. Related methods such as MANOVA and other approaches to multivariate multiple mean comparison are also discussed in Section 3. Some concluding remarks are summarized in the last section.

THE T-SNE PLOT AND ITS ILLUSTRATION

Steps for implementing a t-SNE plot can be summarized as follows.

- Data preprocessing: start by preparing high-dimensional data (e.g., gene expression data). This often involves normalizing or scaling the data so that features have similar scales. Ensure the data is in matrix form, with samples as rows and features (such as gene expression levels) as columns.

- Compute pairwise distances: t-SNE works by comparing the pairwise similarities between data points. For each data point in the high-dimensional space, compute the distances to every other point. This is typically done using Euclidean distance.

- Convert distances to probabilities: for each data point, convert the pairwise distances to probability distributions that represent how likely two points are neighbors in high-dimensional space. This is done using a Gaussian probability distribution, where closer points have higher probabilities of being neighbors.

- Project data into low-dimensional space: t-SNE maps the high-dimensional data into a lower-dimensional space (usually 2D for simple visualization purpose) by finding a similar probability distribution of points in the new space. t-SNE minimizes the difference (or divergence) between the two distributions (highdimensional and low-dimensional) using a cost function called Kullback-Leibler (KL) divergence [8,9].

- Optimize the mapping: t-SNE uses an optimization method, typically gradient descent, to iteratively adjust the points in the lower-dimensional space. The goal is to find a configuration in the lower-dimensional space that best reflects the distances and relationships from the high-dimensional space.

- Visualize the results: once the optimization process is complete, the data points are mapped to a lower-dimensional space. These can be visualized in a scatter plot. The clusters or groupings formed can reveal meaningful patterns or relationships in the data.

- Parameter tuning: t-SNE has parameters such as "perplexity" (which balances local and global data structure) and "learning rate" that can be adjusted to improve the results. These parameters are typically set through trial and error or based on the data characteristics.

There are three key parameters in implementing the t-SNE plot: perplexity, it controls how to balance local and global aspects of the data; learning rate: it affects the speed and effectiveness of the optimization process; and number of iterations: the number of optimization steps, increasing this can improve the quality of the mapping. This implementation is typically done using libraries like Python's `scikit-learn` or `TensorFlow`, which have built-in functions to perform t-SNE. 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. A practical implementation of the t-SNE plot can be done by running the Rtsne package that is free to download and running install. packages("Rtsne") under the R command line after installing the general R software from https://cran.r-project.org/, and then running library(Rtsne) under the R command window. We illustrate its application in the following example.

Example: The gene expression dataset consists of gene mapping data of different time points using mass spectrometry technology. The original data were collected from a study of Metabolic Syndrome (MetS), which is a collection of metabolic risk factors, including central adiposity, hyperglycemia, hypertension and dyslipidemia. Yu ZR, et al. [10] studied the identification of potential biomarkers of metabolic syndrome by taking urine samples from 36 male patients who were policemen in the

City of Tianjin. The original data were obtained through both positive mode and negative mode method by mass spectrometry technology. The experimental data from 36 patients were matched paired with 36 normal patients as the control group. Both the negative-mode and positive-mode data consist of a total of 36 patients in the experimental group and 36 patients in the control group. The urine sample was carried out for 869 time points for negative-mode data and for 552 time points for positive-mode data using mass spectrometry technology. We tale each observation as the gene expression data for 36 patients in the trial group. So the data the dimension p=36, the sample size n=869 for the negative-mode data and n=552 for the positive-mode data. Both of the set of negative-mode data and the set of positive-mode data consist of a total of 36 patients in the trial (experimental) group and 36 patients in the control group. The original data are available upon request.

When determining the possible number of clusters for a gene expression dataset, the elbow plot [11] can be implemented, which is a method of plotting the WCSS=withincluster sum of squares against the number of clusters. The elbow method is based on the k-means algorithm which is provided by the R package "kmeans". The elbow plot for the negative-mode dataset is given in Figure 1.

From the elbow plots, it seems that the better number of clusters for both negative-mode and positive-mode data may be 3, 4, and 5. We summarize the number of observations (the gene expression levels) for each case in Tables 1,2.

The t-SNE plots for both negative-mode and positive-mode data with different number of clusters under different perplexity parameters are shown in the following Figures 2-7.

Summary observation from Figures 1-7: for both negative-mode and positive-mode data, a better classification of the data seems to be four clusters because: 1) classification of three to four clusters leads to a big decrease in the elbow plot as shown in Figure 1; 2) classification of four to five clusters only leads to a small decrease in the elbow plot as shown in Figure 1; and 3) the classification of five clusters shows more overlapped observations than those for the classification of four clusters. Therefore, it is better option to classify both negative-mode and positive-mode data into four clusters, respectively.

Table 1: Classification of negative-mode data into different number of clusters for the trial group.

	No. of observations					
No. of clusters	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Total**
3	3	38	828			869
4	732	21	113	3		869
5	145	14	3	37	670	869

Table 2: Classification of positive-mode data into different number of clusters for the trial group.

	No. of observations					
No. of clusters	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Total**
3	30	517	5			552
4	482	10	56	4		552
5	92	411	10	35	4	552

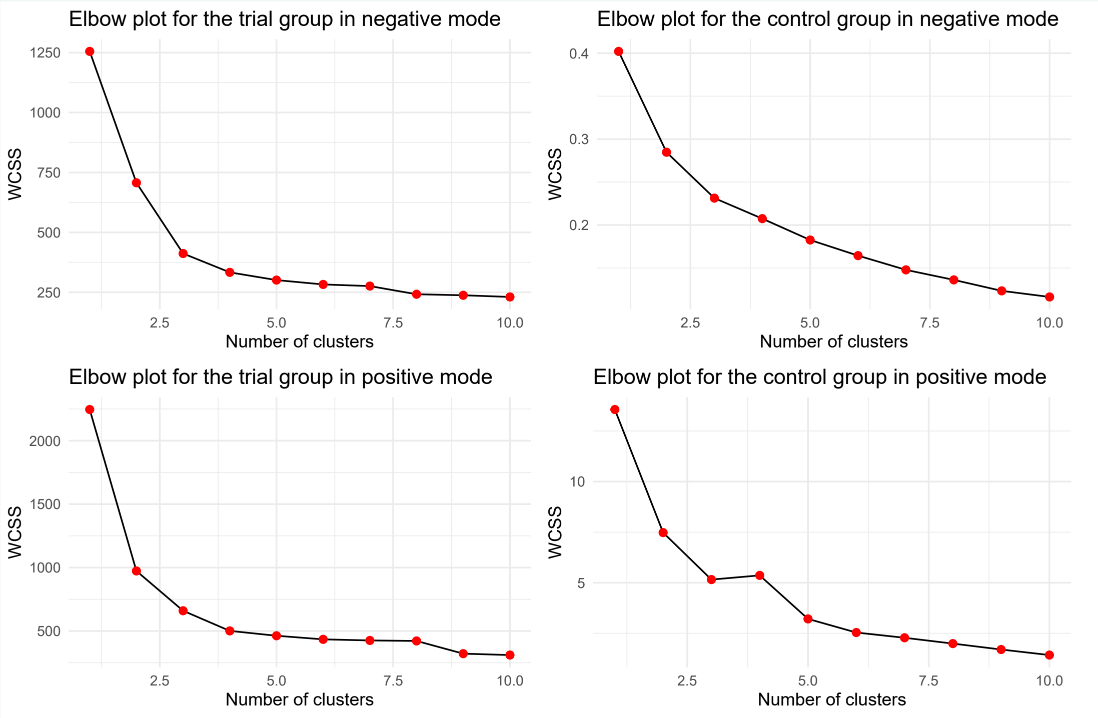


Figure 1 Elbow plots for negative-mode and positive-mode data for the trial and control groups.

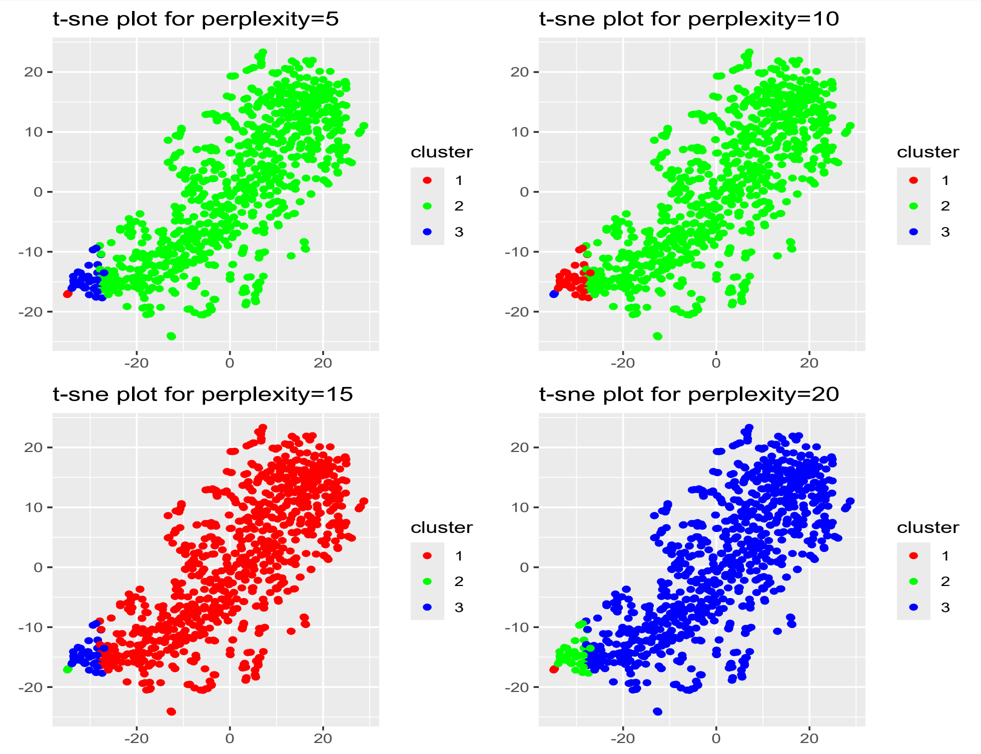


Figure 2 t-SNE plot for negative-mode data from 36 patients in the trial group (clusters=3).

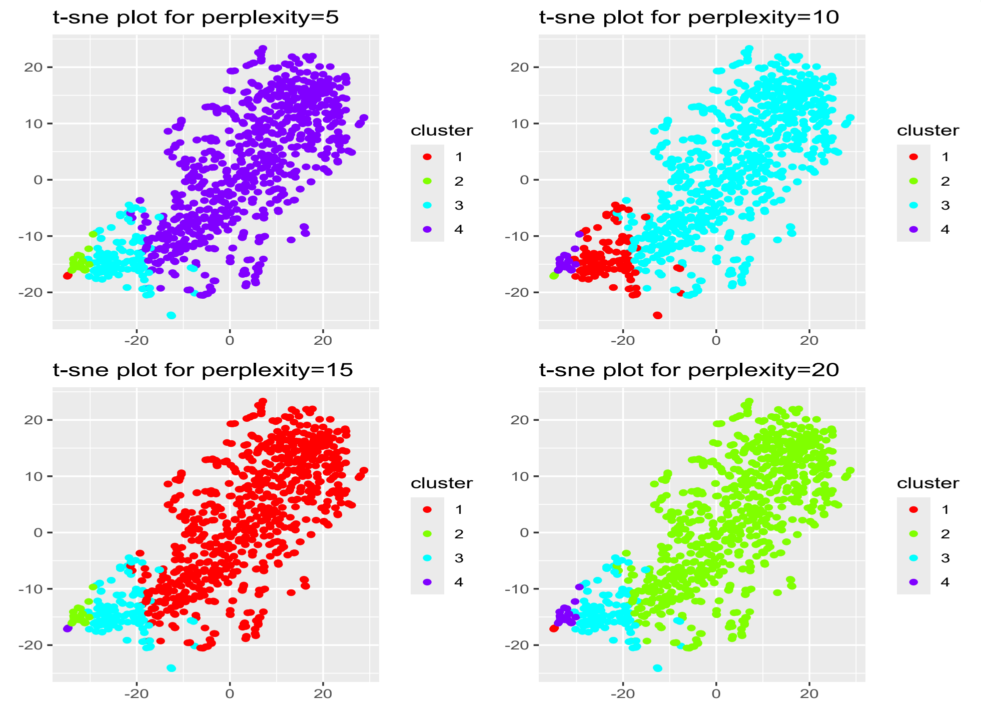


Figure 3 t-SNE plot for negative-mode data from 36 patients in the trial group (clusters=4).

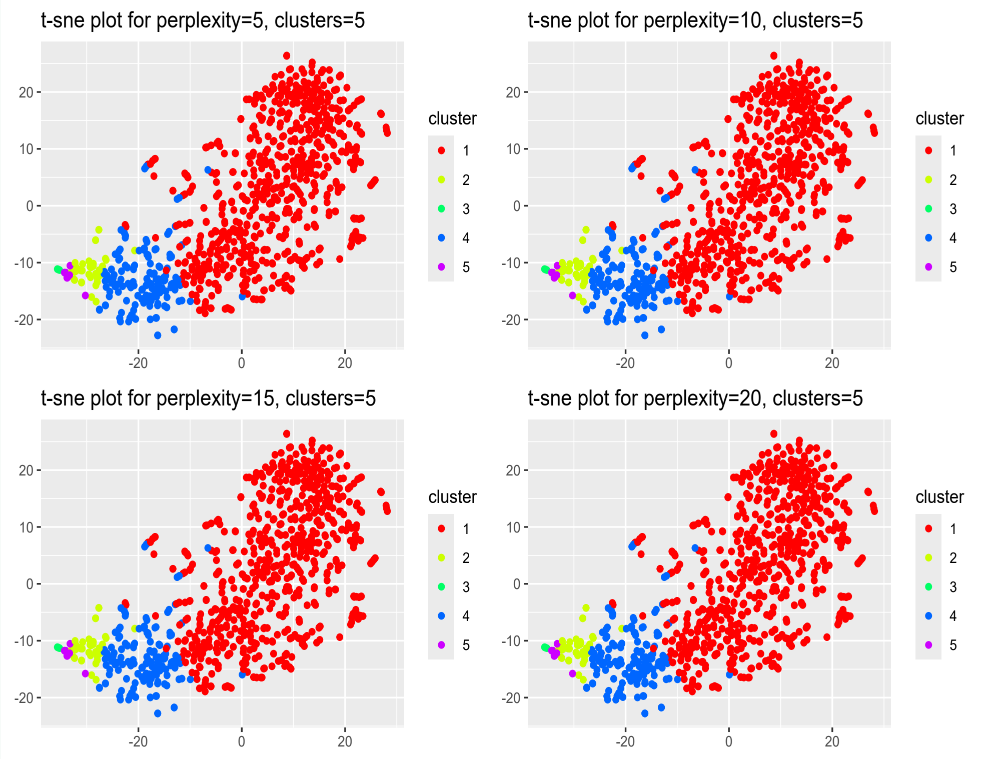


Figure 4 t-SNE plot for negative-mode data from 36 patients in the trial group (clusters=5).

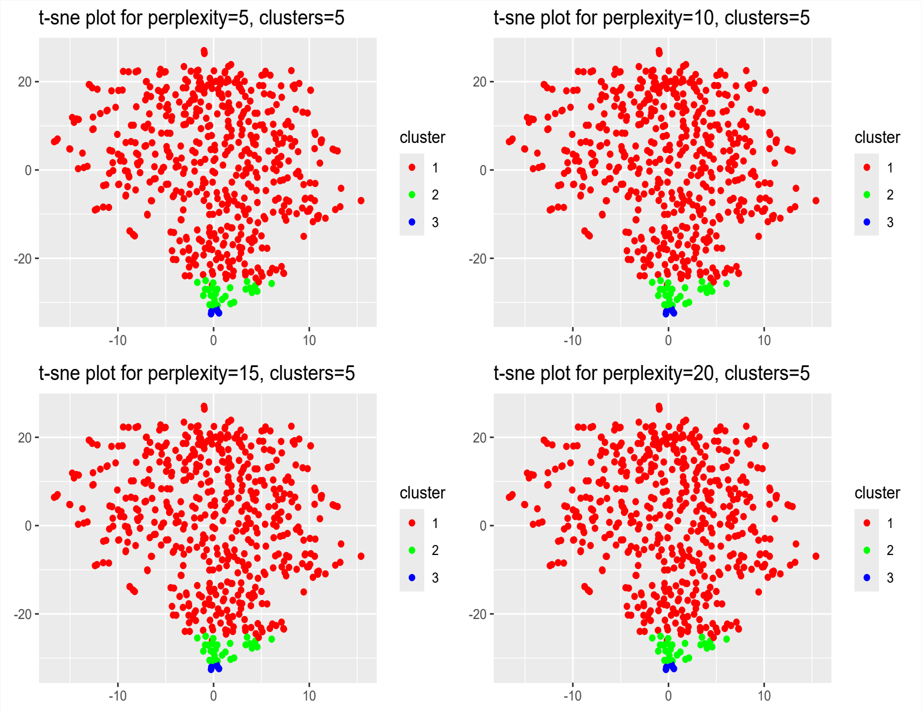
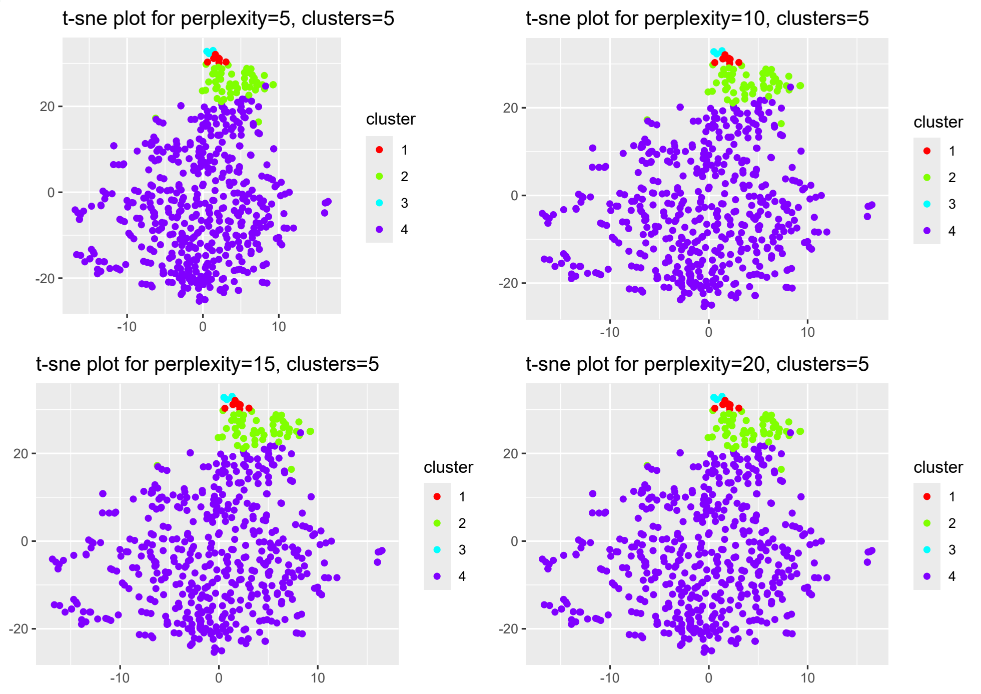


Figure 5 t-SNE plot for positive-mode data from 36 patients in the trial group.



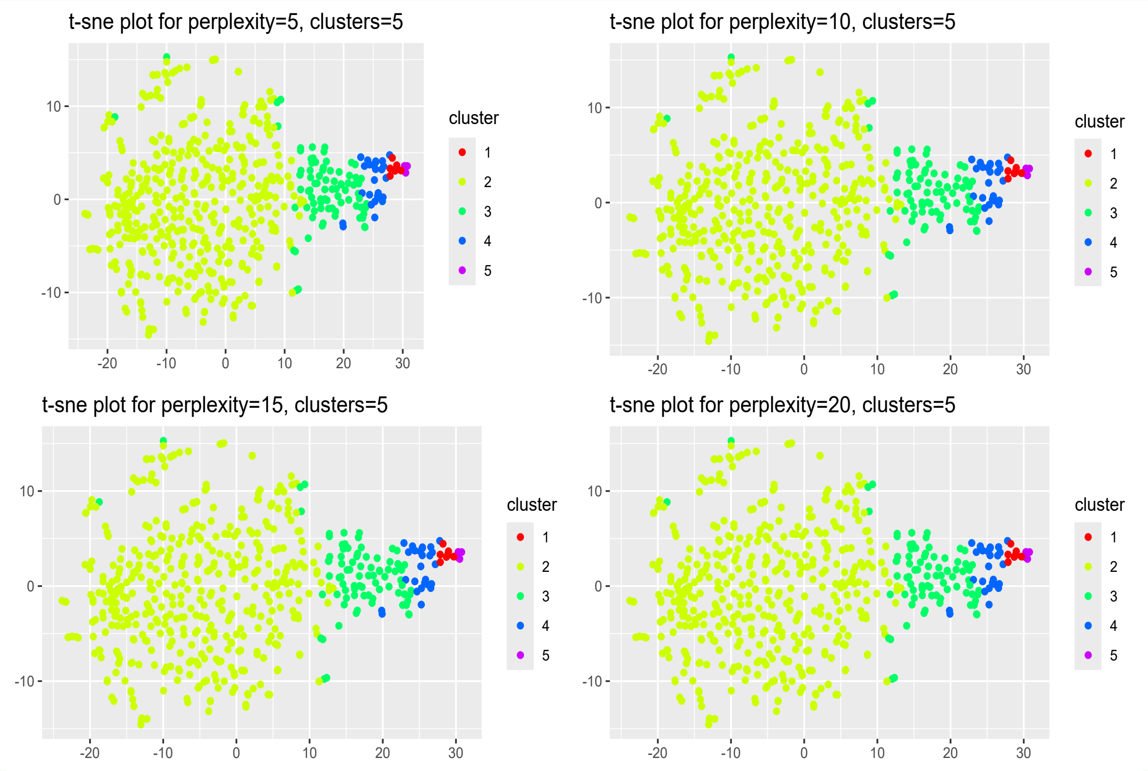
t-sne plot for perplexity=5, clusters=5

t-sne plot for perplexity=10, clusters=5

t-sne plot for perplexity=15, clusters=5

t-sne plot for perplexity=20, clusters=5

Figure 6 t-SNE plot for positive-mode data from 36 patients in the trial group (clusters=4).



t-sne plot for perplexity=5, clusters=5

t-sne plot for perplexity=10, clusters=5

t-sne plot for perplexity=15, clusters=5

t-sne plot for perplexity=20, clusters=5

Figure 7 t-SNE plot for positive-mode data from 36 patients in the trial group (clusters=5).

The generalized F-test for multiple mean comparison

In this section, we will employ the generalized F-test [5] for multiple mean comparison among the four clusters of negative-mode and positive-mode data in the last section. For an i.i.d. (independent identically distributed) sample from each cluster, we follow the notation in [5]. Let

μ! = the mean level of observations in cluster i, $i = 1, \dots, k$, $k \geq 2$

We want to test if the mean levels for all k clusters are the same. That is, we want to test the statistical hypothesis

$$H_0: \mu_1 = \cdots = \mu_k, \qquad (1)$$

against the general alternative hypothesis that at least two means differ. Note that each mean level is a vector of p-dimension with p=the number of patients in each mode (negative or positive)=36. The sample size n=the gene expression levels (n=869 for negative-mode data, and n=552 for positive-mode data). Denote the observations from each cluster by

$$X_i = (x_{i1}, \dots, x_{in_i})': \ n_i \times p, \ \ i = 1, \dots, k; \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}: \ n \times p,$$

$$n = \sum_{i=1}^{k} n_i. \qquad (2)$$

It is known that hypothesis (1) is the problem of Multivariate Analysis of Variance (MANOVA). It can be tested by the classical Wilks-Lambda statistic [1] if the total sample size is greater than the data dimension (i.e., $n = \sum_{i=1}^{k} n_i > p$). Here we employ the generalized F-test [5] to test hypothesis (1) and the two-by-two paired comparison between any two clusters. The generalized F-test is applicable for both cases of $n > p$ and $n \leq p$. According to [5], the generalized F-test is computed as follows. Let

$$Z = AX, \ \ A = (a_{ij}): (n-1) \times n, \ \ a_{ij} = \begin{cases} \dfrac{1}{\sqrt{i(i+1)}}, & j = 1, \dots, i, \\ \dfrac{-i}{\sqrt{i(i+1)}}, & j = i+1, \\ 0, & otherwise \end{cases} \qquad (3)$$

Define the eigenvalue-eigenvector problem

$$\left(\frac{1}{n-1} Z'Z\right) D = D\Lambda, \qquad (4)$$

Where $D = (d_1, \dots, d_p)$: $p \times p$ consists of the eigenvectors, and $\Lambda = diag(\lambda_1, \dots, \lambda_p))$ is a diagonal matrix consists of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Let

$$z_s = Zd_s = (z_{s1}, \dots, z_{n-1})': \ (n-1) \times 1, \ \ s = 1, \dots, q = \min(n-1, p) - 1,$$

$$\bar{z}_s = \frac{1}{n-1} \sum_{i=1}^{n-1} z_{si}, \qquad F_s = \frac{(n-1)\bar{z}_s^{\,2}}{\frac{1}{n-2} \sum_{i=1}^{n-1} (z_{si} - \bar{z}_s)^2}. \qquad (5)$$

The generalized F-statistic is defined by

$$GF(q) = \max_{1 \leq s \leq q} \{F_s\}. \qquad (6)$$

Liang and Tang [5] gave the probability distribution of GF by

$$P(GF(q) < x) \approx F(x; 1, n-2)^q, \qquad (7)$$

under the null hypothesis (1) and the multivariate normal assumption with equal covariance matrices for the data from all clusters, and the sample size is large enough. Here $F(x; 1, n-2)$ stands for the distribution function of the F-distribution $F(1, n-2)$ and $q = \min(n-1, p) - 1$=the number of positive eigenvalues in (4). The p-value of the generalized F-test (6) for hypothesis (1) is computed by

$$pValue \approx 1 - F(GF_0; 1, n-2)^q, \qquad (8)$$

Where GF_0 is computed from the sample data through equations (2)-(6). A small p-value implies rejection of hypothesis (1).

The number q in constructing the generalized F-statistic (6) acts as the projection dimension in the PCA (principal component analysis) problem (4). In general q should be chosen as the number of positive eigenvalues to ensure there exists variation in the data projected to the PCA direction. Based on the PCA theory [12], the idea of explanation of variation of the eigenvalues can be employed to determine q, for example, q can be chosen as the number such that the first q eigenvalues can explain a given percentage $0 < e < 1$:

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq e \quad \text{and} \quad \frac{\sum_{i=1}^{q-1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} < e \qquad (9)$$

We apply the GF-test (6) to the multiple mean comparisons for the clusters in tables 1,2 and summarize the p-values for all comparisons in tables 3,4, where

$$q_1 \text{ satisfies } \frac{\sum_{i=1}^{q_1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq .90 \quad \text{and} \quad \frac{\sum_{i=1}^{q_1-1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} < .90$$

$$q_2 \text{ satisfies } \frac{\sum_{i=1}^{q_2} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq .80 \quad \text{and} \quad \frac{\sum_{i=1}^{q_2-1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} < .80$$

$$q_3 \text{ satisfies } \frac{\sum_{i=1}^{q_3} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq .70 \quad \text{and} \quad \frac{\sum_{i=1}^{q_3-1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} < .70$$

$$q_4 \text{ satisfies } \frac{\sum_{i=1}^{q_4} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq .60 \quad \text{and} \quad \frac{\sum_{i=1}^{q_4-1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} < .60$$

That is, q_1 is the smallest projection dimension such that the first q_1 PCA directions can explain a least 90% data variation; q, is the smallest projection dimension such that the first q, PCA directions can explain a least 80% data variation; $q/$ is the smallest projection dimension such that the first $q/$ PCA directions can explain a least 70% data variation; and $q0$ is the smallest projection dimension such that the first $q0$ PCA directions can explain a least 60% data variation (Tables 3,4).

The p-values in tables 3,4 imply that it is practicable to classify the negative-mode data for the trial group into either 3, 4, or 5 clusters depending on practical needs, while it is impractical to

Table 3: P-values for the generalized F-tests for multiple mean comparison (negative-mode data).

Proj. dim.	$q = 35$	$q_{\#} = 5$	$q_{,} = 2$	$q_{/} = 1$	$q_{0} = 1$
No. of clusters 3 p-value	$GF(q)$ $= 376$	$GF(q_{\#})$ $= 376$	$GF(q_{,})$ $= 376$	$GF(q_{/})$ $= 376$	$GF(q_{0})$ $= 376$
	0.0000	0.0000	0.0000	0.0000	0.0000
No. of clusters 4 p-value	$GF(q) = 54$	$GF(q_{\#}) = 54$	$GF(q_{,}) = 54$	$GF(q_{/}) = 54$	$GF(q_{0}) = 54$
	0.0000	0.0000	0.0000	0.0000	0.0000
No. of clusters 5 p-value	$GF(q) = 79$	$GF(q_{\#}) = 79$	$GF(q_{,}) = 79$	$GF(q_{/}) = 79$	$GF(q_{0}) = 79$
	0.0000	0.0000	0.0000	0.0000	0.0000

Abbreviations: Proj. dim.: Projection Dimension

Table 4: P-values for the generalized F-tests for multiple mean comparison (positive-mode data).

Proj. dim.	$q = 35$	$q_{\#} = 3$	$q_{,} = 1$	$q_{/} = 1$	$q_{0} = 1$
No. of clusters 3 p-value	$GF(q) = 28$	$GF(q_{\#}) = 28$	$GF(q_{,})= 28$	$GF(q_{/})= 28$	$GF(q_{0})= 28$
	0.0000	0.0000	0.0000	0.0000	0.0000
No. of clusters 4 p-value	$GF(q) = 30$	$GF(q_{\#})= 30$	$GF(q_{,})= 30$	$GF(q_{/})= 30$	$GF(q_{0})= 30$
	0.0000	0.0000	0.0000	0.0000	0.0000
No. of clusters 5 p-value	$GF(q) = 19$	$GF(q_{\#})= 3.7$	$GF(q_{,})= 3.7$	$GF(q_{/})= 3.7$	$GF(q_{0})= 3.7$
	0.0004	0.1596	0.0563	0.0563	0.0563

Abbreviations: Proj. dim.: Projection Dimension

classify the positive-mode data for the trial group into 5 clusters. This conclusion is basically consistent with the data visualization from the t-SNE plots in Figures 2-7. Note that when choosing the projection dimension q=1, the generalized F-test reduces to the Läuter's exact F-test [13] under the null hypothesis (1) and the multivariate normal assumption on all clustered data. This makes generalized F-test more applicable than the classical MANOVA method, which generally relies on large sample size to obtain the asymptotic null distribution.

CONCLUDING REMARKS

In conclusion, the integration of the t-SNE plot with the generalized F-test offers a powerful and flexible approach to analyzing gene expression data. By utilizing the t-SNE plot, this method effectively tackles the challenge of classifying high-dimensional gene expression data into distinct clusters, leveraging nonlinear dimensionality reduction to reveal patterns and relationships that may be missed by linear methods. The visual clarity of the classified data enhances understanding and interpretation, making it a valuable tool in biological data analysis.

The application of the generalized F-test for multiple mean comparison in this context is particularly innovative. It addresses the limitations of traditional MANOVA method, which may fail when the dimensionality of the data exceeds the total sample sizes of the clusters. By employing Principal Component Analysis (PCA) to project the data into various dimensions, the test can be applied across different projected spaces. This layered approach not only strengthens the robustness of the hypothesis testing but also offers a means of validating results across multiple dimensions, adding another layer of reliability.

The combination of the t-SNE plot and the generalized F-test bridges visualization and statistical rigor, allowing researchers to make data-driven decisions with confidence. However, the rejection of the null hypothesis in this context opens new doors for further investigation. Future research should focus on the underlying biological significance of the gene clusters and explore other statistical methodologies for deeper insight into differential gene expression patterns. Overall, this approach represents a meaningful step forward in the exploration and understanding of high-dimensional gene expression data, promising to contribute significantly to the field of bioinformatics.

ACKNOWLEDGMENTS

The authors thank Prof. Xi Chen at Tianjin Medical University, China, for providing the real gene expression data in the example of this paper.

Conflicts of Interest: There is no conflict of interest.

REFERENCES

1. Anderson TW. An Introduction to Multivariate Statistical Analysis, 3rd ed. Wiley Inc: USA; 2003.
2. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000; 290: 2323-2326.
3. Tenenbaum JB, Silva VD, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000; 290: 2319-2323.
4. van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008; 9: 2579-2605.
5. Liang J, Tang ML. Generalized F-tests for the multivariate normal mean. Computational Statistics & Data Analysis. 2009; 53: 1177-1190.
6. Konstorum A, Jekel N, Vidal E, Laubenbacher R. Comparative analysis of linear and nonlinear dimension reduction techniques on mass cytometry data. BioRxi. 2018.
7. Amir el-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. 2013; 31: 545-552.
8. Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics. 1951; 22: 79-86.
9. Joyce JM. Kullback-leibler divergence. In: International Encyclopedia of Statistical Science. Lovric M, editor. Springer; 2014. pp. 720-722.
10. Yu ZR, Ning Y, Yu H, Tang NJ. A HPLC-Q-TOF-MS-based urinary metabolomic approach to identification of potential biomarkers of metabolic syndrome. J Huazhong Univ Sci Technolog Med Sci. 2014; 34: 276-283.
11. Ketchen DJ, Shook CL. The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal 1996; 17: 441-458.
12. Jolliffe IT. Principal Component Analysis. Springer, New York, 1986.
13. Läuter J. Exact t and F tests for analyzing studies with multiple endpoints. Biometrics 1996; 52: 964-930.