frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 27 April 2021
doi: 10.3389/fmolb.2021.643014

Microbiome Data Analysis by Symmetric Non-negative Matrix Factorization With Local and Global Regularization

Junmin Zhao[1†], Yuanyuan Ma[2†] and Lifang Liu[3]*

[1] School of Computer and Data Science, Henan University of Urban Construction, Pingdingshan, China, [2] School of Computer and Information Engineering, Anyang Normal University, Anyang, China, [3] School of Education, Anyang Normal University, Anyang, China

A network is an efficient tool to organize complicated data. The Laplacian graph has attracted more and more attention for its good properties and has been applied to many tasks including clustering, feature selection, and so on. Recently, studies have indicated that though the Laplacian graph can capture the global information of data, it lacks the power to capture fine-grained structure inherent in network. In contrast, a Vicus matrix can make full use of local topological information from the data. Given this consideration, in this paper we simultaneously introduce Laplacian and Vicus graphs into a symmetric non-negative matrix factorization framework (LVSNMF) to seek and exploit the global and local structure patterns that inherent in the original data. Extensive experiments are conducted on three real datasets (cancer, cell populations, and microbiome data). The experimental results show the proposed LVSNMF algorithm significantly outperforms other competing algorithms, suggesting its potential in biological data analysis.

Keywords: matrix factorization, Laplacian regularization, Vicus graph, microbiome, local structure

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Ma Ying Jun,
Central China Normal University, China
Xingpeng Jiang,
Central China Normal University, China

****Correspondence:***
Yuanyuan Ma
chonghua_1983@126.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

Received: *17 December 2020*
Accepted: *22 February 2021*
Published: *27 April 2021*

Citation:
Zhao J, Ma Y and Liu L (2021) Microbiome Data Analysis by Symmetric Non-negative Matrix Factorization With Local and Global Regularization. Front. Mol. Biosci. 8:643014. doi: 10.3389/fmolb.2021.643014

INTRODUCTION

With the development of high-throughput metagenomic sequencing and 16S sequencing technologies, more and more biological data have been accumulated. Generally, these sequences have very complicated characteristics, making discoveries and identification of latent relations among samples very daunting. In order to reach a good understanding of the roles that the microbiome plays in the health and disease states of humans, many plans, including the Human Microbiome Plan (HMP) (Turnbaugh et al., 2007), integrative Human Microbiome Plan (iHMP) (The Integrative Hmp (iHMP) Research Network Consortium, 2019), and the Metagenomics of Human Intestinal Tract (MetaHIT) (Qin et al., 2010), have been launched. These actions pave the way for researchers to further explore the complex relationships residing in microbiome data.

Arumugam et al. classified the microbiome into different enterotypes and pointed out the significance of a functional analysis to reveal the interactions among microorganisms (Arumugam et al., 2011; Siezen and Kleerebezem, 2011). Clustering approaches and similarity measurements were applied to elucidate the influence that various factors impose on the identification of enterotypes (Koren et al., 2013). Jiang et al. (2012) proposed a new approach based on non-negative matrix factorization (NMF) to identify the structure and functions of complex microbial

communities across environmental samples. Wu et al. (2016) developed a stable NMF method, staNMF, and obtained a novel and concise representation of spatial gene expression patterns. As a clustering technology, NMF has attracted a lot of attention in terms of its good data representation capabilities. In NMF, samples or features can be viewed as the linear additive combination of basis vectors. Meanwhile, the membership label of each sample can be assigned by the corresponding coefficient matrix. When the data have a linear structure, NMF usually achieves better performance. However, in the real world, data points are generally embedded in a non-linear manifold; thus, adopting other ways (such as graphs) to describe latent relationships among data points is a better choice (Kuang et al., 2012). Symmetric non-negative matrix factorization (SNMF) takes a similarity matrix as input and outputs a cluster indicator. In this process, a similarity matrix can be obtained in many ways, such as through a Gaussian kernel, polynomial kernel, linear kernel, and so on. Kuang et al. (2012, 2015) designed an effective SNMF algorithm to model the complex relationships contained in non-linear data that outperforms many NMF-based approaches. Ma et al. (2016a) developed HSNMF, a method that combined SNMF with a second-order graph to explore microbiome data.

Recently, the graph regularization framework has been successfully applied in many fields including bioinformatics, image processing, and text mining, achieving good performance. Cai et al. (2010) proposed the GNMF algorithm to reveal the potential patterns of several datasets. Specifically, GNMF used Laplacian regularization (Lr) to encode the intrinsic geometrical structure presented in the original data. Subsequently, a number of variants based on Lr were generated (He et al., 2015; Ma et al., 2016b, 2017). Although these methods obtained some interesting findings, they may ignore some important aspects. For example, a traditional Laplacian graph captures the global structure of a data matrix, which is insufficient in biology research, where local topologies need to be sought and utilized effectively (Wang et al., 2017). Moreover, recently emerging methods developed to capture local topological information have been shown to obviously outperform global algorithms (Roweis and Saul, 2000; Wu and Schölkopf, 2007; Jiang and Hu, 2014). These methods aim to reconstruct each data point using its local neighbors and are shown to be robust and insensitive to outliers. Vicus, an alternative local spectral matrix, can effectively capture the local geometrical information from neighboring nodes to model biological interactions, and it shares many similar properties with Laplacian matrices; for example, both matrices are symmetric and positive semidefinite (psd), and both have non-negative, real-valued eigenvalues (Wang et al., 2017). Compared with Laplacian graphs, Vicus graphs, which are constructed via local subnetworks, are more robust with respect to noise and can lessen the influence of outliers to some extent. In this paper, we first used similarity graphs to establish the complicated relationships in samples; second, given these graphs, we constructed Laplacian and Vicus spectral matrices; finally, we integrated the Vicus (and/or Laplacian) matrix into an SNMF framework to conduct downstream analysis, such as clustering, visualization, and so on.

In view of the above considerations, in this paper we introduce Laplacian and Vicus matrices (Wang et al., 2017) to simultaneously model the global and local structure connections residing within the data and compare their performance with the methods only based on Laplacian or Vicus graphs on several real datasets. Our experiments include tumor classification, microbiome samples identification and so on. The experimental results show that the proposed algorithm outperforms other baseline and competing approaches, which demonstrates its efficiency and effectiveness in microbiome data analysis. **Figure 1** gives an illustrative example.

The contribution of this work lies in the fact that (1) an effective clustering algorithm has been proposed and can be easily expanded to other applications and (2) to our knowledge, this is the first attempt to integrate global and local structure information into an SNMF framework to conduct microbiome data analysis. The rest of this paper is organized as follows: in the next section a brief statement of SNMF is given. Then Lr, the formulation of Vicus, and the proposed algorithm are also provided. In section "Results and Discussion", extensive experiments are conducted, and the experimental results and comparisons analysis are presented. Section "Conclusion" summarizes the conclusions and further research plans.

MATERIALS AND METHODS

Symmetric Non-negative Matrix Factorization

In SNMF, given an $n \times n$ symmetric matrix A and a reduced rank k, SNMF seeks to find the best factorization so that $A = HH^T$, where H can be viewed as the cluster indicator. The objective of SNMF can be formalized as follows:

$$O = \min_{H \geq 0} \left\| A - H H^T \right\|_F^2 \tag{1}$$

Where $A \in R_+^{n \times n}$ with $A = A^T$, $H \in R_+^{n \times k}$, and $\|\bullet\|$ denotes the c of a matrix. Compared with NMF, SNMF concerns only the factorized similarity matrix A and doesn't consider whether the structure of the data is linear or non-linear. Once A is given, SNMF conducts factorization similar to that of NMF. Therefore, SNMF is more suitable to modeling unknown data. For a matrix A, A_{ij}, the ij-th element of A, denotes the similarity score between the ith and the jth data points. Similarity metrics can take many forms. One common way is to use the Gaussian kernel function to construct a weighted similarity matrix:

$$w_{ij} = \exp\left[-\frac{\|X_i - X_j\|_F^2}{\sigma_i \sigma_j} \right] \ (i \neq j) \tag{2}$$

where, X_i denotes the ith data point (sample), σ_i is the Euclidean distance between X_i and its d-th neighborhood. We set $d = 7$ as suggested in the literature of Zelnik-Manor and Perona (2005). It is noted that the diagonal elements of similarity W are set to be zeros to eliminate self-similarity. Next, we only retain those edges linking nodes with their p nearest neighbors $N(i)$. Thus,

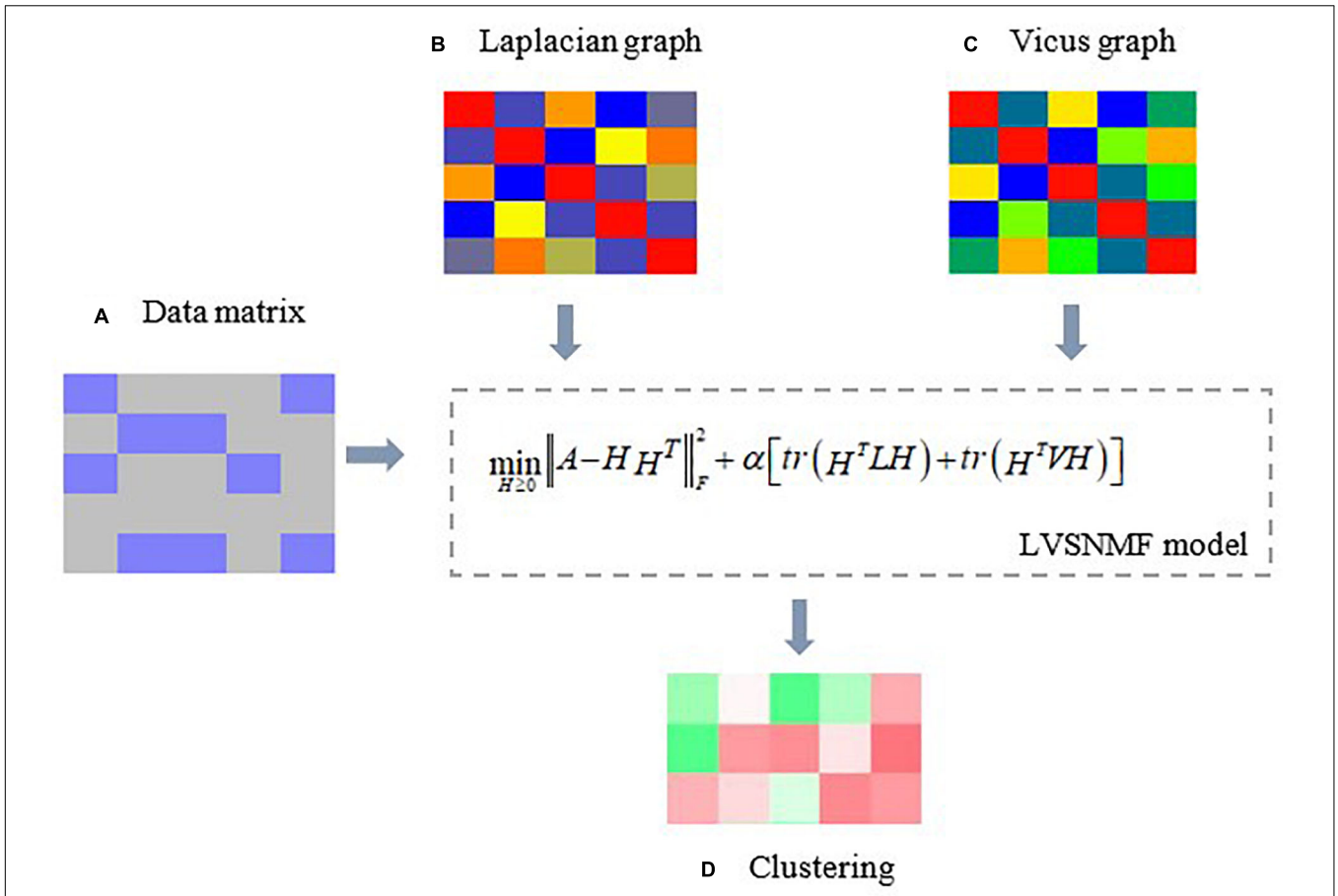

FIGURE 1 | An illustrative example of the proposed LVSNMF algorithm. **(A)** the original data matrix (gene expression matrix, microbiome abundance profile matrix, and so on. **(B)** Laplacian graph used to maintain manifold consistence assumptions. **(C)** A Vicus graph explores the local geometrical structure in the data. Then c and d are introduced into the proposed LVSNMF model, which integrated the global (Laplacian) and local (Vicus) geometrical structure of the original data. **(D)** the clustering result given by LVSNMF.

the weighted matrix W derived from Equation 2 can be rewritten as:

$$w_{ij} = \begin{cases} w_{ij}, & if \quad i \in N(j) \ or \ j \in N(i); \\ 0, & otherwise. \end{cases} \tag{3}$$

Where $N(i)$ is the set of neighbors of the ith node.

As suggested in [20], W can be transformed into the normalized form:

$$A = D^{-1/2} W D^{-1/2} \tag{4}$$

here, $D_{ii} = \sum_{j=1}^{n} w_{ij}$ denotes the degree matrix.

Finally, by multiplication update rules SNMF can obtain the locally optimal solution:

$$h_{ik} \leftarrow h_{ik} \frac{(AH)_{ik}}{\left(HH^T H\right)_{ik}} \tag{5}$$

Laplacian Graph

Given P, a weighted matrix, let L be the Laplacian matrix; L can be defined as:

$$L = D - P \tag{6}$$

where D is a degree matrix and $D_{ii} = \sum_{j=1}^{n} d_{ij}$. The normalized cut version of L can be formalized as:

$$L = I - D^{-1/2} P D^{-1/2} \tag{7}$$

where I denotes the identity matrix. Note that the general Laplacian matrix (Equation 6) is used in the complete experiments, and the normalized version of L is used to be input for spectral methods. In the process of constructing a weighted graph P, many similarity functions can be used, such as the inner product function, kernel function, and so on. In our experiments, the Gaussian kernel function (Equation 2) is employed to establish P.

Vicus Graph

As described in Wang et al. (2017), Vicus graphs have the same properties as Laplacian graphs. For example, both matrices are symmetric and positive semidefinite, and the eigenvector corresponding to the smallest eigenvalue 0 of both matrices is the constant vector **1**. Compared with the Laplacian graph, the Vicus graph can capture the local structure within the data. In this

subsection, we demonstrate the process of constructing a Vicus graph in detail. Note that the Vicus matrix is constructed based on a sample-sample similarity matrix, which can be computed via any similarity function, such as the Gaussian kernel, the cosine kernel functions, and so on. In this paper, we used the Gaussian kernel for computing the similarities between any two samples. The differences between the Laplacian and the Vicus graphs lies in the fact that the former describes the global structure information inherent in the data whereas the latter captures well the fine-grained topological information present in the biological network. The intuition is that we can use the local connection information from neighboring nodes to make the Vicus matrix more robust with respect to noise. Thus, it helps to lessen the influence of outliers.

Let $\{x_1, x_2, \cdots, x_n\}$ be a set of data points. Then v_i corresponding to x_i denotes the ith vertex in a weighted network P, and $N(i)$ represents x_i's neighbors, not including x_i. Here, the neighborhood size of all nodes is consistent ($\|N_i\| = k, i = 1, 2, \cdots, n$).

The main assumption behind Vicus is that the cluster label of the ith data point can be inferred from its nearest neighbors $N(i)$. First, a subnetwork $P_i = (V_i, E_i)$ is extracted such that $V_i = N(i) \bigcup x_i$ and E_i represents the edges connecting all points inherent in V_i. Using the label diffusion algorithm (Zhou et al., 2004), a virtual label indicator vector $c^k_{v_i}$ can be reconstructed as:

$$c^k_{v_i} = (1-\alpha)\,(I - \alpha\, S_i)^{-1}\, q^k_{v_i}, \quad 1 \le k \le C \tag{8}$$

Where $\alpha \in (01)$ is a constant. In all our experiments, α is set 0.9 as suggested in Wang et al. (2017). C is the number of clusters. $q^k_{v_i}$ is the scaled cluster indicator of $P_i.S_i$ denoting the normalized transition matrix, i.e., $S_i(u,t) = P_i(u,t) \Big/ \sum_{l=1}^{K+1} P_i(u,l)$. $c^k_{v_i}$ is a vector of $K+1$ elements; here $\overline{q}^k_i = c^k_{v_i}[K+1]$ is the estimate of how likely it is that node i belongs to the kth cluster. The goal is maximal concordance between $\overline{q}^k_i$ and q^k_i. Let $\beta_i \in R^{K+1}$ be the ith row of the matrix $(1-\alpha)\,(I - \alpha\, S_i)^{-1}$, representing label propagation at its terminal state. We set $\overline{q}^k_i = \beta_i\, q^k_{v_i}$. Thus, $\overline{q}^k_i$ can be approximated by

$$\overline{q}^k_i \approx \frac{\beta_i[1:K]\, q^k_{N(i)}}{1 - \beta_i[K+1]} \tag{9}$$

Where $\beta_i[1:K]$ is the first K elements of β_i and $\beta_i[K+1]$ denotes the $(K+1)$th element in β_i .

Next, we use matrix B to represent the linear relationship $\overline{q}^k \approx B\, q^k, \quad k = 1, 2, \cdots, C$:

$$B_{ij} = \begin{cases} \frac{\beta_i[j]}{1-\beta_i[K+1]} & \text{if } x_j \in N(i) \text{ and } x_j \text{ is the } j\text{-th element in } N(i); \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

In order to minimize the difference between $\overline{q}^k$ and q^k, we can adopt a simple objective as listed below:

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{k=1}^{C}\left(\overline{q}^k_i - q^k_i\right)^2 = \sum_{k=1}^{C}\left\|\overline{q}^k - q^k\right\|^2 \approx \sum_{k=1}^{C}\left\|q^k - B\, q^k\right\|^2 \\ &= Tr\left(Q^T\,(I-B)^T\,(I-B)\,Q\right) \end{aligned} \tag{11}$$

Here, $\mathrm{Tr}(\bullet)$ denotes the trace of a matrix. Setting $V = (I-B)^T\,(I-B)$, we thus obtain the Vicus matrix. Similarly to the traditional spectral clustering formulation, by performing eigen-decomposition of V we can obtain clustering results. In this paper, we use V as constraint term to preserve the local manifold structure of the data.

Note that in order to better reveal the complicated relationships among microbiome samples, we simultaneously introduce the Laplacian and Vicus spectral matrices into the objective function of SNMF. The details are given below, in section "Symmetric Non-negative Matrix Factorization Based on Laplacian and Vicus Regularization."

Symmetric Non-negative Matrix Factorization Based on Laplacian and Vicus Regularization

Based on analysis above, we combine Laplacian and Vicus matrices into the Symmetric Non-negative Factorization framework to explore the global and local structure inherent in the data. The proposed algorithm, namely LVSNMF, takes full advantage of the global and local consistency of the data to model complex relationships between different samples. The final objective can be defined simply as:

$$O = \min_{H \ge 0} \left\|A - H\,H^T\right\|_F^2 + \alpha\left[tr\left(H^T\,LH\right) + tr\left(H^T\,VH\right)\right] \tag{12}$$

Where α is a regularization parameter and used to balance the trade-off between matrix reconstruction errors and spatial structure preservation. The second term in Equation 12 simultaneously takes into account the global information (Laplacian) and local structure (Vicus) of the data.

To minimize the objective of LVSNMF, we use multiplicative update rules to solve the optimal problem. The updating formula of H can be obtained as follows:

$$h_{ik} \leftarrow h_{ik}\, \frac{(AH)_{ik} + \alpha\left[\left(P + V^-\right)H\right]_{ik}}{\left(H\,H^T\,H\right)_{ik} + \alpha\left[\left(D + V^+\right)H\right]_{ik}} \tag{13}$$

Here, $V = V^+ - V^-$.

LVSNMF integrates the global and local similarity information inherent in the data, and it can therefore obtain better performance than just using the Laplacian graph or other methods based on local diffusion information from neighboring nodes.

Evaluation Metrics

The proposed LVSNMF and competing algorithms are evaluated by comparing the generated labels of all samples with the ground truth contained in the datasets. Two common cluster metrics, accuracy (AC) and normalized mutual information (NMI), are used to evaluate the performance of the proposed LVSNMF algorithm. Generally, the higher AC and NMI values achieved, the better the clustering quality is. More detailed information on these two metrics can be found in Xu et al. (2003).

Datasets

Three datasets are used in our experiments. The first one is a cancer dataset from TCGA, the second is a pollen dataset, and the last is a human microbiome dataset from HMP[1]. The important statistics of these three datasets are summarized in **Table 1**.

Lung cancer: This is a benchmark dataset including five cancer subtypes. It can be downloaded from https://jundongl.github.io/scikit-feature/datasets.html.

Pollen: these data consist of 11 cell populations, containing neural cells and blood cells. These data are obtained directly from Pollen et al. (2014).

Human microbiome data: these data consist of 637 samples drawn from seven body sites: one gut (stool), one vagina (posterior fornix), one nasal (anterior nares), one skin site (retroauricular crease), and three oral sites (supragingival plaque, tongue dorsum, and buccal mucosa). Each sample consists of 710 microorganisms. The relative abundance of each at species level wase estimated by MetaphlAn. All the data can be downloaded from the HMP website.

RESULTS AND DISCUSSION

Experimental Results

In this section, we conduct extensive experiments on these three datasets. The experimental results are shown in **Table 2**. From this table, we can see that LVSNMF outperforms the second best algorithm at 0.49/1.93% points in terms of AC/NMI on the Lung dataset, 3.22/2.17% points on the Pollen dataset, and the 2.09/0.61% points on HMP dataset.

The results in **Table 2** are obtained when $p = 12$ and $k = 15$ across all three datasets. Here, p denotes the number of neighbors in constructing similarity matrix A, k is the number of local neighbors in constructing a Vicus matrix. For other values, LVSNMF still outperforms these competing algorithms in most cases. Note that in our experiments NNDSVD (Boutsidis and Gallopoulos, 2008) is utilized to enhance the initiation of SNMF-based algorithms, which result in rapid reduction of reconstruction errors.

[1]http://hmpdacc.org/

TABLE 1 | Statistics of the two datasets.

Dataset	Number of samples	Number of features	Number of clusters
Lung cancer	203	3,312	5
Pollen	249	14,805	11
HMP	637	710	7

TABLE 2 | The best performance in three real datasets.

	Accuracy (%)			Normalized mutual information (%)		
	Lung	Pollen	HMP	Lung	Pollen	HMP
SNMF	83.74	84.66	87.84	67.51	86.49	84.52
SNMF + Laplacian	90.64	85.94	88.27	70.03	87.33	84.46
SNMF + Vicus	90.15	85.60	88.49	71.26	87.12	84.95
LVSNMF	91.13	89.16	90.58	72.96	89.50	85.56

Parameter Analysis

In the proposed LVSNMF method, there is one essential parameter: the regularization parameter α. In fact, Laplacian and Vicus matrices should have different weight parameters. In this study, we set them to be equal for convenience. Parameter α reflects the extent to which we want to exert punishment for violating the manifold consistency hypothesis.

In our experiments, the values of α are set to be in the range (0.001 0.005 0.01 0.05 0.1 0.5 1 10). **Table 2** reports the best performance of all algorithms on three datasets.

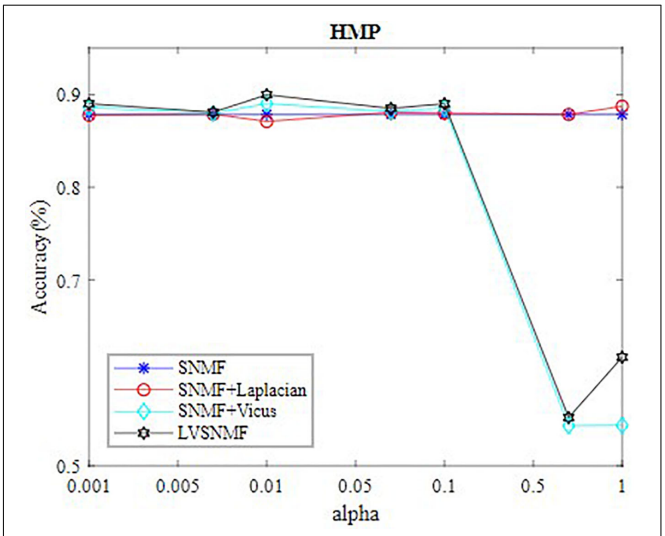

FIGURE 2 | The best performance of LVSNMF as αincreases.

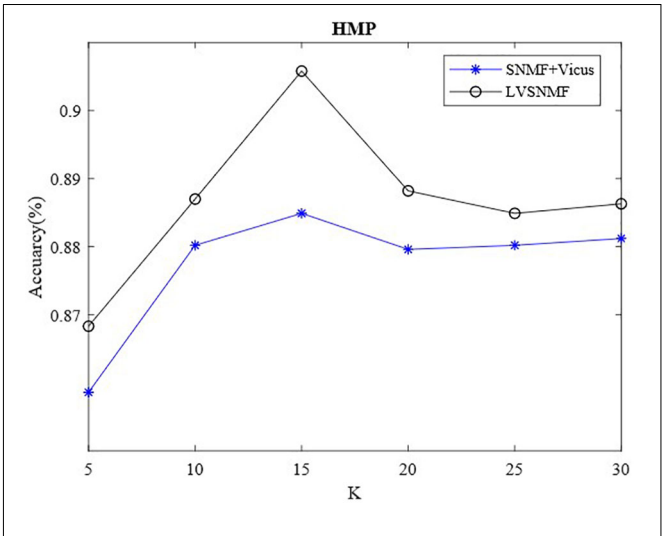

FIGURE 3 | The performance of LVSNMF varies as K increases.

In order to intuitively observe performance of LVSNMF, we draw curve of LVSNMF vs.α on HMP dataset. **Figure 2** gives the performance curve of LVSNMF as α increases.

Comparison and Discussion

In order to further demonstrate the effectiveness of LVSNMF, we compare the performance of LVSNMF with three other algorithms on the HMP dataset when α varies in the range (0.001 0.005 0.01 0.05 0.1 0.5 1). As shown in **Figure 1**, in the interval (0.001, 0.1), LVSNMF achieves consistently good performance as α increases. When α, LVSNMF obtains the best performance (90.58%/85.56% in terms of AC/NMI), and performs slightly better than the second best algorithm. One possible reason is that samples from HMP are noiseless and the Laplacian has successfully captured sufficient structure information such that the performance of LVSNMF might not be obvious.

On the other hand, on the pollen dataset the proposed LVSNMF algorithm significantly outperforms other competing algorithms (as **Table 2** shows). This suggests that LVSNMF has the ability to seek and find the latent structure inherent in the data.

Vicus has an important hyperparameter, the number of neighbors *K*. To validate the influence of *K* on the performance of LVSNMF, we also conduct additional experiments on the HMP dataset. **Figure 3** shows that the performance of LVSNMF varies with K ($\alpha = 0.01$).

As shown in **Figure 3**, two algorithms based on Vicus regularization have consistently good performance as *K* varies within the interval (10, 30), especially when *K* equals 15; then they obtain the best performance. This suggests that the proposed LVSNMF is robust with respect to the number of neighbors *K*. Unlike the SNMF + Vicus algorithm, LVSNMF also takes into account the global information of the data; therefore, it can achieve the better performance in most cases.

In summary, the proposed LVSNMF method can adequately capture the global and local structure of the data. The experimental results on three real datasets demonstrate its efficiency and effectiveness.

CONCLUSION

In this paper, we propose a novel approach, called LVSNMF, to conduct microbiome data analysis. In LVSNMF, the global and local structure similarities are encoded in Laplacian and Vicus matrices, respectively. Extensive experiments are executed on three datasets, and the results show that the proposed LVSNMF algorithm significantly outperforms other baseline or state-of-the-art methods, which demonstrate its efficiency and effectiveness on microbiome data analysis.

Although LVSNMF achieves good performance, it concerns only sample clustering. In the real world, the relationship among microbes is often subtle and complicated. Therefore, modeling microbial interactions is important to dissect the mechanism behind diseases related to the microbiome. In the future, we will develop new methods to construct microbial interaction networks and seek appropriate ways to describe the functional or genetic similarities among microbes, such as phylogenetic trees, metabolic abilities, and so on.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/chonghua-1983/LVSNMF.

AUTHOR CONTRIBUTIONS

JZ and YM wrote the manuscript, designed the algorithms, and conducted all the experiments. YM developed the concept for the structure and content of the manuscript. LL critically revised the final manuscript. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work is supported by the Key Research Projects of Henan Higher Education Institutions (No. 20B520002) and The Key Technology R&D Program of Henan Province (No. 202102310561).

ACKNOWLEDGMENTS

This manuscript is recommended by the 5th Computational Bioinformatics Conference.

REFERENCES

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180.

Boutsidis, C., and Gallopoulos, E. (2008). SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recogn.* 41, 1350–1362. doi: 10.1016/j.patcog.2007.09.010

Cai, D., He, X., Han, J., and Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/tpami.2010.231

He, Y.-C., Lu, H.-T., Huang, L., and Shi, X.-H. (2015). Non-negative matrix factorization with pairwise constraints and graph laplacian. *Neural Process. Lett.* 42, 167–185. doi: 10.1007/s11063-014-9350-0

Jiang, X., and Hu, X. (2014). Inferring microbial interaction networks based on consensus similarity network fusion. *Sci. China Life Sci.* 57, 1115–1120. doi: 10.1007/s11427-014-4735-x

Jiang, X., Weitz, J. S., and Dushoff, J. (2012). A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data. *J. Math. Biol.* 64, 697–711. doi: 10.1007/s00285-011-0428-2

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., et al. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* 9:e1002863. doi: 10.1371/journal.pcbi.1002863

Kuang, D., Ding, C., and Park, H. (2012). "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the 2012 SIAM*

International Conference on Data Mining: SIAM, (Philadelphia, PA: Society for Industrial and Applied Mathematics), 106–117.

Kuang, D., Yun, S., and Park, H. (2015). SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. *J. Global Optim.* 62, 545–574. doi: 10.1007/s10898-014-0247-2

Ma, Y., Hu, X., He, T., and Jiang, X. (2016a). Hessian regularization based symmetric nonnegative matrix factorization for clustering gene expression and microbiome data. *Methods* 111, 80–84. doi: 10.1016/j.ymeth.2016.06.017

Ma, Y., Hu, X., He, T., and Jiang, X. (2016b). "Multi-view clustering microbiome data by joint symmetric nonnegative matrix factorization with Laplacian regularization," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (New York, NY: IEEE), 625–630.

Ma, Y., Hu, X., He, T., and Jiang, X. (2017). Clustering and integrating of heterogeneous microbiome data by joint symmetric nonnegative matrix factorization with laplacian regularization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 788–795 doi: 10.1109/tcbb.2017.2756628

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32:1053. doi: 10.1038/nbt.2967

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65.

Roweis, S. T., and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326. doi: 10.1126/science.290.5500.2323

Siezen, R. J., and Kleerebezem, M. (2011). The human gut microbiome: are we our enterotypes? *Microbial Biotechnol.* 4, 550–553. doi: 10.1111/j.1751-7915.2011.00290.x

The Integrative Hmp (iHMP) Research Network Consortium (2019). The integrative human microbiome project. *Nature* 569, 641–648. doi: 10.1038/s41586-019-1238-8

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810.

Wang, B., Huang, L., Zhu, Y., Kundaje, A., Batzoglou, S., and Goldenberg, A. (2017). Vicus: exploiting local structures to improve network-based analysis of biological data. *PLoS Comput. Biol.* 13:e1005621. doi: 10.1371/journal.pcbi.1005621

Wu, S., Joseph, A., Hammonds, A. S., Celniker, S. E., Yu, B., and Frise, E. (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc. Natl. Acad. Sci. U S A.* 113, 4290–4295. doi: 10.1073/pnas.1521171113

Wu, M., and Schölkopf, B. (2007). A local learning approach for clustering. *Adv. Neural Inform. Process. Systems* 19, 1529–1536.

Xu, W., Liu, X., and Gong, Y. (2003). "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, (New York, NY: ACM), 267–273.

Zelnik-Manor, L., and Perona, P. (2005). Self-tuning spectral clustering. *Adv. Neural Inform. Process. Systems* 17, 1601–1608.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. *Adv. Neural Inform. Process. Systems* 16, 321–328.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.