Survey on High-Dimensional Medical Data Clustering

Velmurugan Arresh Balaji
Chonnam National University
Gwangju, South Korea
arreshvnass@gmail.com

Chulwoong Choi
Chonnam National University
Gwangju, South Korea
sentilemon02@gmail.com

Kyungbaek Kim
Chonnam National University
Gwangju, South Korea
kyungbaekkim@jun.ac.kr

ABSTRACT

In a relative less span of time we can process and store a large quantity of data due to technological advancements. There is a rapid change in the nature of data, specifically, the dimensional property of data, mostly in multi and high-dimensional. In terms of heterogeneity of data, Data analysis have becoming a humungous task, Because the volume and complexity in data has been increasing incrementally. In data mining, there is a tool called Data clustering, used in many disciplines in order to extract the meaningful knowledge from seemingly unstructured data. The high-dimensional patient's health records such as immune system status, DICOM Images like CT/PET images, electronic medical records, microarray data like gene expressions, genetic background, etc., In this article we have done a survey on high dimensional medical data clustering and different approaches related to this problem. It also focusses on the real-life applications and recent methods in high dimensional cluster analysis.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; **Unsupervised learning**;

KEYWORDS

High Dimensional Data, Clinical data, LLRR Clustering, Gaussian mixture copulas, Pathway based deep clustering

ACM Reference Format:
Velmurugan Arresh Balaji, Chulwoong Choi, and Kyungbaek Kim. 2020. Survey on High-Dimensional Medical Data Clustering. In *The 9th International Conference on Smart Media and Applications (SMA 2020), September 17–19, 2020, Jeju, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3426020.3426071

1 INTRODUCTION

In terms of clustering, the high-dimensional data, a kind of data which has more attributes and it is identified to have certain challenges in clustering. In 1954, the data clustering came into the picture, when an article had this term in its title which dealt with the data such as anthropological data. The term Cluster analysis has its origin from various fields like data mining, statistics, machine learning, artificial intelligence, biology, etc, and they used variant names for analyzing the cluster. Some of them are as follows: Q Analysis, data visualization, typology, numerical taxonomy, clumping, and so on. To classify the data objects, there is a famous data mining technique called clustering. And the data objects classification is based on dissimilarity and similarity among them.

In fields such as texture segmentation, data compression, vector quantization and computer vision, the most common form of data mining technique is clustering. Clustering methods have combinatorial nature, as the problem scale increases, these methods will be computationally intractable. For machine Learning techniques, Electronic medical records such as high dimensional data and more complexed data provides opportunities for clustering approaches. To overcome the curse of Dimensionality and to obtain efficient processing time, the cluster analysis depends on the dimension reduction. Even though, for cluster analysis and dimension reduction, a variety of approaches available, but there is no direct approach to identify the best combination technique from both families to obtain the desired result. An in-depth understanding about the raw data, analytical process, configuring parameters, and intermediary results are required, to derive precise and efficient insights from the high dimensional medical datasets.

Pathway-based clustering methods have been developed by incorporating biological pathway databases. Pathway-based analysis plays an important role in understanding collective biological functions of genes and their impact on the phenotypic changes of the patients [31]. Pathifier discovered several pathways which are significantly associated with patients' survivals in glioblastoma and colorectal cancer [7]. The method inferred pathway deregulation scores from gene expression data and then performed clustering. R-PathCluster identified two subtypes of glioblastoma and several pathways associated with the cancer progression [19]. In the study, pathway scores were generated from gene expression and subtypes were identified by clustering the pathway scores.

On gene-expression datasets, mixture models were found to outperform widely used classical methods like K-means and hierarchical clustering [29]. Mixture models are a principled statistical approach to clustering, where inferred clusters can be interpreted through the lens of the underlying dimensional assumptions. Although mixture models are over-parameterized in high dimensions, which make parameter inference difficult, variable selection techniques and parsimonious covariance structures alleviate the problems to a large extent and enable their use in subtyping [1] [32] [27]. However, through the choice of the multivariate distribution, model-based clustering imposes distributional assumptions on the marginals, along each dimension, and these marginal distributions are assumed or forced to be identical (e.g. a multivariate normal

SMA 2020, September 17–19, 2020, Jeju, Republic of Korea

ACM ISBN 978-1-4503-8925-9/20/09…$15.00
https://doi.org/10.1145/3426020.3426071

imposes univariate normal distribution on each marginal); such assumptions restrict their modeling flexibility. For many applications, including clustering, a semi-parametric model works well, where copulas are used to model dependency patterns, assuming no fixed parametric model for the marginals. However, copulas are rarely used in high dimensional settings either parameter estimation is intractable, or they lose their modeling flexibility [13].

One of the important clustering methods is Subspace clustering, and its purpose is to divide data into different clusters. It relies on the principle that data within a cluster have the similar subspace representation. When cluster corresponds to a subspace, and the data from the same subspace can be grouped into a class. Hence, Studying the data subspace would be the way to overcome this Subspace clustering challenge. For subspace clustering a method called Low-Rank Representation (LRR) [17] [16] has been proposed and has received more and more attention for its success in learning the underlying low-dimensional subspace structure.

The data variance function would be the simplest evaluation for feature selection, when compared to the conventional evaluation functions. For data recovery and to find features, filter methods like the Principal Component Anlaysis (PCA) method and its variants can be used. Since there is no clear reason, the effective discrimination between data points in different classes, can't be confirmed with the features selected. The Laplacian Score (LS) method was proposed for selecting features with high identification. By comparing with other methods, LS method is independent, and it is also an 'filter' method [11]. The LS method constructs a nearest neighbour graph for preserving the local geometric structure. The data space's local structure can be reflected by the selected features. The global data structure's influences have been ignored by the LS method and it concentrate more relationship among the local data points. This might be a limitation for the features selected from the given multiple subspaces data, hence, the discrimination effects of the selected features might be reduced. For a multiple subspace dataset, it is hard to represent and characterize the global data structures with satisfaction, using the feature selection method. By potential lowdimensional subspaces, representing the high-dimensional data is the key to the LLR method [33]. In the bioinformatics field, LRR has achieved great success in gene expression data mining. For example, to identify the subspace gene clusters, the LRR method and obtained good results [6]. The Summary of various High Dimensional Medical data clustering Techniques was given in Table.1.

2 METHODS OF HIGH DIMENSIONAL MEDICAL DATA CLUSTERING

2.1 Pathway based deep clustering

Due to the non-linear relationship between the patient's survivals and genomic data in cancer, the conventional similarity/distance between data-based clustering approaches failed to cluster. Due to the high nonlinearity, it was reported that even binary classification for survival prediction (short-long-term survival prediction) produces a low Area Under the Curve (AUC) of around 0.65 in a balanced dataset. It may be caused by multiple intermediate complex biological processes between genomic data and survivals. Hence, the pathway-based clustering methods were developed to assimilate biological pathway databases, to understand the collective

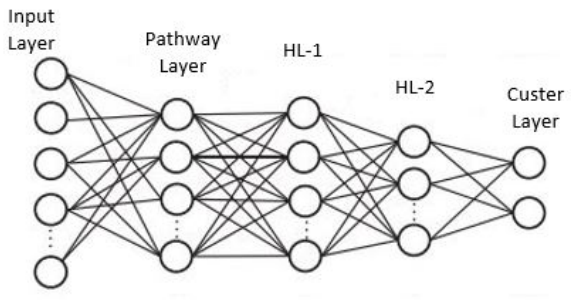

Figure 1: A Simple Pathway based Model.

gene biological function and their effects of phenotypic changes in patients. The analysis based on pathway plays an important role. Usually, the cancer dataset consists of clinical features and distinct molecular features for multiple subtypes of cancer. For cancer treatment, the response will be different for different cancer subtypes, so the cancer sub-tying helps in making a decision and will improve the personalized treatment. For the identification of molecular subtypes, enormous genomic data related to cancer is currently available all across the world. To identify the cancer subtypes that are unique in both clinical and genetic, many unsupervised machine learning approaches have been applied on the molecular data of tumour samples so far. Due to the challenging non-linearity and high-throughput genomic data, most of the clustering methods were failed to cluster the patients effectively. A pathway based deep clustering method (PACL) [19] for molecular subtyping of cancer, which incorporates gene expression and biological pathway database to group patients into cancer subtypes.

By learning complex nonlinear effects and hierarchical of pathways, they developed a model to find the high-level representations of biological data. And they compared the performance of pathway-based deep clustering model with a number of benchmark clustering methods that recently have been proposed in cancer subtypes. With the help of log-rank tests, they assessed the hypothesis, whether the different survivals are associated with the clusters (subtypes) or not. PACL showed the lowest p-value in the log-rank test. The PACL method got very low p-value and outperformed other benchmark methods. It demonstrates the patient groups clustered by PACL may correspond to subtypes which are significantly associated with distinct survival distributions. In biological pathway level, the PACL model can helps to identify the subtypes and interpret them effectively. In the Figure 1, the hierarchical associations and nonlinear of pathways to a cluster are performed by the two hidden layers HL-1 and HL-2. And it also indicates the activity states that are associated with multiple pathways like active/inactive. Finally, the hidden layers capture the multiple pathway's group effects and does not represent the biological processes explicitly.

2.2 Gaussian mixture copulas

The identification of sub-populations of patients with similar characteristics, called patient subtyping and it is important for realizing the goals of precision medicine. Accurate subtyping is crucial for tailoring therapeutic strategies that can potentially lead to reduced mortality and morbidity. Model-based clustering, such as Gaussian

Technique used	Algorithm	Description	Author	Title	Year
Low rank Representation	LLLR	This algorithm uses Linearized Adaptive Method with Adaptive Penalty (LADMAP)	Juan wang et.al	Laplacian Regularized low-rank Representation	2019
	NSLRG-S	The Rand Index (RI) is used to compute how similar the result of clustering is to the actual classification.	Lu.C et.al	Clustering based on score function	2020
Gaussian Mixture Copulas	HD-GMCM	In HD-GMCM, uses a LASSO penaltry approach for the mean parameters.	Siva Rajesh et.al	Dependancy based Sub-typing	2020
Molecular Sub-typing	PACL	PACL effectively clusters high-dimensional gene expression data, which are nonlinearly associated to patients' survivals.	Tejaswini et.al	Pathway-based deep Clustering	2020

Figure 2: Table 1: Summary of various High Dimensional Medical data clustering Techniques.

mixture models, provides a principled and interpretable methodology that is widely used to identify subtypes. However, they impose identical marginal distributions on each variable; such assumptions restrict their modeling flexibility and deteriorates clustering performance. The statistical framework of copulas provides a modular parameterization of multivariate distributions that decouples the modeling of marginals from the dependencies between them. This allows each marginal to be chosen independently from any distribution and the dependency model offers a richer characterization than single-number metrics like Pearson's or Spearman's correlation coefficients. Thus, when interest lies mainly in discovering feature dependencies, copulas provide an elegant model of dependencies with no restrictive assumptions on the marginals. Such models have been used extensively in finance and more recently in dependency clustering that discovers clusters based on their dependency patterns [26].

The HD-GMCM model [12] used the statistical framework of copulas to decouple the modeling of marginals from the dependencies between them. Current copula-based methods cannot scale to high dimensions due to challenges in parameter inference. The HD-GMCM model addresses these challenges, and to our knowledge, it is the first copula-based clustering method that can fit high-dimensional data. Empirically, such modeling not only uncovers latent structure and also leads to better clustering and meaningful clinical subtypes in terms of patient's survival rate. They reviewed about model-based clustering of high-dimensional data and discussed the information loss due to dimensionality reduction before clustering[4]. Two categories of approaches have been developed for high-dimensional model-based data clustering.

2.2.1 Subspace clustering methods. The Subspace clustering method seeks to reduce the dimensionality of the cluster in local and clusters the data in the same time. Mixture of factor analyzers [21] is one of such conventional approach. Along with the data dimensionality, there will be quadratic growth of the number of covariance parameters, and constrained covariance structures were introduced in MFA through a family of parsimonious Gaussian mixture models (PGMMs)[22] [23]. A High Dimensional Data clustering [5] uses a combination of subspace clustering and parsimonious modeling for Gaussian mixture models. To determine the data cluster structure, the relevant variables will be selected by the variable selection methods for clustering. The variable selection methods in model based was reviewed [9]. A broad class of techniques uses penalized clustering criteria [25]. A clustering method (VarSelLCM) [20], with an efficient inference algorithm through the use of a new information criterion. Using this criterion, it simplifies model selection and works particularly well for $p > n$ cases, for moderately large n. For subtyping, Gaussian graphical models were used for high-dimensional clustering [27]. That adapts to the cluster's scale, sample size and no. of clusters using a penalized likelihood.

2.2.2 MCopulas and mixture models. In various articles [10] [14][26] the Mixture of Copulas model approach had been implemented due to the multivariate distribution's flexibility characterization. But they failed to address the high-dimensional data clustering problem. Then, in bivariate copulas hierarchical collection's, for model estimation and selection, there is The Vine Copulas that scale the high dimensions, but it performs moderate at an exponentially increasing cost and complexity [24]. A discussion about fitting on high dimensional data with copulas and provided a comparative study about copulas with machine learning models [8]. The GMCM model [28], which is different from Mixture of copulas, it belongs to the family of copulas where the Gaussian mixture model followed by a copula density(latent).The copulabased clustering is more advantageous because, it deviates the dependency need for GMCM parameter estimation as the clusters can be directly inferred.

For parameter estimation of GMCM, an expectation maximization (EM) algorithm [2] was designed. Later, to fit both real and ordinal data for clustering and they also designed a mixed and improved algorithm called Expectation Maximum and Gibbs sampling-based approach. The paper [3] discussed computational and statistical hurdles in GMCM parameter estimation and offer some resolutions, but none of these methods work well for clustering high-dimensional data. In a related work, the paper [15] studied a specific case of GMCM to examine the consistency and reliability of experiments with high throughputs and designed a reproducibility analysis (META-ANALYSIS) method. On real high-dimensional

gene-expression and clinical datasets, the HDGMCM had outperformed the state-of-the-art model-based clustering techniques, by virtue of modeling non-Gaussian data and being robust to outliers through the use of Gaussian mixture copulas. The Clusters obtained from HD-GMCM can be interpreted based on the dependencies of the model, that offers a new way of characterizing subtypes.

2.3 LLRR Clustering Method

In Genomic data clustering, the Laplacian regularized Low-Rank Representation (LLRR) clustering method [30], clusters the cancer samples from the high dimensional genomic data. The LLRR method, roughly treats the samples extracted from many low-rank subspaces combinations as the high dimensional genomic data. Based on a dictionary, The LLRR method gets the lowest-rank representation matrix. Because a manifold based Laplacian regularization is introduced into LLRR. Besides capturing the global geometric structure, the LLRR can capture the intrinsic local structure of high-dimensional observation data better than the LRR method. In addition, the original data themselves are selected as a dictionary, so the lowest rank represented by the LLRR method is actually a similar expression between the samples. Therefore, corresponding to the LRR matrix, the high similarity samples are considered to come from the same subspace and are grouped into a class. After decomposition of LRR we can obtain the Low-rank matrix and the subspace clustering is also based on that matrix as well. The data's subspace structure can be preserved well by this LRR method. In numerous research areas, many approaches based on LRR has been applied and implemented. To detect differentially expressed genes by using Discriminative information and joint Graph Laplacian (GLD-RNMF) an effective non-negative matrix factorization can be used. For reducing the noise an LRR method with graph regularization were developed. And for clustering subspace, a graph regularized LRR (LRRGR)is also developed. To differentiate the expressed genes an LRR method was proposed. And they proposed an LRR with Mixed-norm Laplacian regularization (MLLRR). Within data, the inherent geometric structure was completely considered by these algorithms, in many areas such as selecting features, noise reduction and clustering or segmenting subspaces, they achieved great performance results.

The K-means clustering selects cluster centers randomly, so there will be a small change in the clustering results every time. And such differences affect the clustering methods evaluation as well. For K-means method, for reducing these differences, repeat the experiment on the experimental datasets for 30 times and for the clustering accuracy, the mean value is taken as the performance results. Since the K-means algorithm is used to evaluate the final clustering results, in these experiments, firstly, the experimental data is decomposed by one of these methods, then K-means is repeated 30 times based on a matrix after decomposition and the mean of 30 clustering results is taken as the clustering accuracy. By comparing the experiment results on real genomic data with LRR and MLLRR, it illustrates that In the cancer samples clustering, the LLRR method is more robust to noise and it achieved remarkable performance and it also has an better learning ability about the data's inherent subspace structure.

2.4 NSLRG Based on Score Function

In Bioinformatics, for cancer research, cancer sample clustering is a prominent research area. In depth, testing approaches to select the characteristics of genes from high dimensional data such as gene expressions. A cancer clustering integrated framework called as the non-negative symmetric low-rank representation with graph regularization based on score function (NSLRG-S) [18]. They used high dimensional dataset obtained from The Cancer Genome Atlas (TCGA). By comparing with the similar clustering approaches, the NSLRG-S approach is more efficient. At first, This NSLRG method performs under the NSLR matrix and graph regularization constraints. It is the lowest rank matrix to satisfactorily represent the gene expression data and can capture the global structures and local geometric structures of the raw data. Non-negativity is more consistent with biological modelling. The lowest rank matrix's interpretability was improved by the symmetric constraint. To learn the structure of the gene expression data, the lowest rank matrix was facilitated by both non-negativity and symmetry constraints.

Second, for cancer sample clustering, a lowest rank matrixbased score function, was proposed for selecting feature genes. The genes which are selected, have strong discriminability for realizing different samples classification, then finally a novel framework for feature selection, called as NSLRG-S, which is designed for selecting the feature genes and evaluating them for clustering the cancer samples. Based on this framework, only lower level dimensionality has been achieved for the selected result of the gene expression dataset. In multi-cancer sample clustering, by using the selected result as the experimental data, to find the subsets, this NSLRG-S method had achieved a high recognition rate. In the NSLRG method, based on low-rank matrix, by using score function the feature genes are obtained. The low-rank matrix preserves the raw data's local and global structure. In multi-subspace clustering, it is observed that there is a strong discrimination in the selected genes, when the low-rank matrix is further processed by Score function. For the the raw gene expression dataset, the NSLRG-S method simultaneously considers the data's global and local structure.

For clustering the cancer samples, on the subsection Gene Expression Datasets, the NSLRG-S method was applied. Typically, for addressing a high-dimensional and a small sample size problem, gene expression data mining can be recognized. The applied methods must suffer and from what is known as the dimensionality's curse. This scenario happens because, the more the dimensions contained in the data, the more the unstable result. Therefore, by running the experiment 50 times, it improved the reasonableness of the result. The clustering result's mean is taken as the measurement. In comparison, the clustering results of other methods and NSLRG-S, in most datasets the results of the NSLRGS method outperformed all other methods. In subspace clustering, the features selected from the genes have a high recognition rate. Performance of cancer samples can be significantly improved by the NSLRG-S framework.

3 DISCUSSION

The ineffectiveness of traditional algorithms caused by various factors like general increase in dimensionality and increasing complexity of different computational problems was collectively termed

as the 'curse of dimensionality'. Amid other effects of curse of dimensionality, As the number of dimensions increases, a meaningful differentiation loss is observed between dissimilar and similar objects. New clustering techniques are required, because, high dimensional objects mostly seems to be similar. The recent researches about high-dimensional medical data mainly focused on developing clustering algorithms and techniques. So, the remaining research issues are still open. For automatic data grouping, there is a data mining tool called mutual similarity-based clustering. Each cluster groups objects that are similar to one another, whereas dissimilar objects are assigned to different clusters, possibly separating out noise. In this manner, clusters describe the data structure in an unsupervised manner, i.e., without the need for class labels. For cluster detection, there has been many existing different algorithmic approaches and cluster models. All approaches should require an underlying assessment of similarity between data objects.

4 CONCLUSIONS

In the current Situation, when the appropriate dataset is obtained, many people arises the question that which similar measure should be used. Considering the situation given, one should consider in what way we have to implement our domain knowledge with respect to the situation. Finally, to analyze the given dataset, we have to select a suitable approach. In this modern era, we cannot solve the problems with only one clustering algorithm. Which means that there is no best way itself., It depends purely on the problem we have. To process clinical high-dimensional and largescale datasets, we have the capability by the advancement in Computational ability and technology. This low-rank matrix can preserve. This survey provides a brief review on high-dimensional and large medical datasets and their clustering process, recent trends and approaches and finally the challenges faced by those methods. Moreover, it describes the clustering's fundamental concepts, such as validation and tendency of clustering, and their clustering process in detail. Mainly, this survey was done to present about the medical high-dimensional medical data clustering algorithms and techniques.

5 FUTURE WORK

To develop a Self-Organizing Subspace Clustering model for NonSmall Cell Lung Cancer (NSCLC) high dimensional and Multiview data. Since, it is normally more difficult to work with due to its large number of features or dimensions and to improve the performance of the existing prognosis models for the survival predictions of NSCLC patients.

ACKNOWLEDGMENTS

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (NRF2019M3E5D1A02067961). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B40 12559).

REFERENCES

[1] Jangsun Baek and Geoffrey J McLachlan. 2011. Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics* 27, 9 (2011), 1269–1276.
[2] Sakyajit Bhattacharya and Vaibhav Rajan. 2014. Unsupervised learning using Gaussian mixture copula model. In *21st International Conference on Computational Statistics. Geneva, Switzerland.*
[3] Anders E Bilgrau, Poul S Eriksen, Jakob G Rasmussen, Hans E Johnsen, Karen Dybkær, Martin Bøgsted, et al. 2016. GMCM: Unsupervised clustering and meta-analysis using gaussian mixture copula models. *Journal of Statistical Software* 70, 2 (2016), 1–23.
[4] Charles Bouveyron and Camille Brunet-Saumard. 2014. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* 71 (2014), 52–78.
[5] Charles Bouveyron, Stéphane Girard, and Cordelia Schmid. 2007. High-dimensional data clustering. *Computational Statistics & Data Analysis* 52, 1 (2007), 502–519.
[6] Yan Cui, Chun-Hou Zheng, and Jian Yang. 2013. Identifying subspace gene clusters from microarray data using low-rank representation. *PloS one* 8, 3 (2013), e59377.
[7] Yotam Drier, Michal Sheffer, and Eytan Domany. 2013. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences* 110, 16 (2013), 6388–6393.
[8] Gal Elidan. 2013. Copulas in machine learning. In *Copulae in mathematical and quantitative finance*. Springer, 39–60.
[9] Michael Fop, Thomas Brendan Murphy, et al. 2018. Variable selection methods for model-based clustering. *Statistics Surveys* 12 (2018), 18–65.
[10] Ryohei Fujimaki, Yasuhiro Sogawa, and Satoshi Morinaga. 2011. Online heterogeneous mixture modeling with marginal and copula selection. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 645–653.
[11] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. 2005. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence* 27, 3 (2005), 328–340.
[12] Siva Rajesh Kasa, Sakyajit Bhattacharya, and Vaibhav Rajan. 2020. Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping. *Bioinformatics* 36, 2 (2020), 621–628.
[13] Abbas Khalili and Jiahua Chen. 2007. Variable selection in finite mixture of regression models. *Journal of the american Statistical association* 102, 479 (2007), 1025–1038.
[14] Ioannis Kosmidis and Dimitris Karlis. 2016. Model-based clustering using copulas with applications. *Statistics and computing* 26, 5 (2016), 1079–1099.
[15] Qunhua Li, James B Brown, Haiyan Huang, Peter J Bickel, et al. 2011. Measuring reproducibility of high-throughput experiments. *The annals of applied statistics* 5, 3 (2011), 1752–1779.
[16] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. 2012. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 171–184.
[17] Guangcan Liu, Zhouchen Lin, and Yong Yu. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 663–670.
[18] Conghai Lu, Juan Wang, Jinxing Liu, Chunhou Zheng, Xiangzhen Kong, and Xiaofeng Zhang. 2019. Non-Negative Symmetric Low-Rank Representation Graph Regularized Method for Cancer Clustering Based on Score Function. *Frontiers in Genetics* 10 (2019).
[19] Tejaswini Mallavarapu, Youngsoon Kim, Jung Hun Oh, and Mingon Kang. 2017. R-pathcluster: Identifying cancer subtype of glioblastoma multiforme using pathway-based restricted boltzmann machine. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1183–1188.
[20] Matthieu Marbac and Mohammed Sedki. 2017. Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing* 27, 4 (2017), 1049–1063.
[21] Geoffrey J McLachlan, David Peel, and RW Bean. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* 41, 3-4 (2003), 379–388.
[22] Paul David Mcnicholas and Thomas Brendan Murphy. 2008. Parsimonious Gaussian mixture models. *Statistics and Computing* 18, 3 (2008), 285–296.
[23] Paul David McNicholas, Thomas Brendan Murphy, Aaron F McDaid, and Dermot Frost. 2010. Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis* 54, 3 (2010), 711–723.
[24] Dominik Müller and Claudia Czado. 2018. Representing sparse Gaussian DAGs as sparse R-vines allowing for non-Gaussian dependence. *Journal of Computational and Graphical Statistics* 27, 2 (2018), 334–344.
[25] Wei Pan and Xiaotong Shen. 2007. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* 8, May (2007), 1145–1164.

[26] Mélanie Rey and Volker Roth. 2012. Copula mixture model for dependency-seeking clustering. *arXiv preprint arXiv:1206.6433* (2012).
[27] Nicolas Städler, Frank Dondelinger, Steven M Hill, Rehan Akbani, Yiling Lu, Gordon B Mills, and Sach Mukherjee. 2017. Molecular heterogeneity at the network level: high-dimensional testing, clustering and a TCGA case study. *Bioinformatics* 33, 18 (2017), 2890–2896.
[28] Ashutosh Tewari, Michael J Giering, and Arvind Raghunathan. 2011. Parametric characterization of multimodal distributions with non-gaussian modes. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 286–292.
[29] Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng, and George C Tseng. 2006. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* 22, 19 (2006), 2405–2412.
[30] Juan Wang, Jin-Xing Liu, Xiang-Zhen Kong, Sha-Sha Yuan, and Ling-Yun Dai. 2019. Laplacian regularized low-rank representation for cancer samples clustering. *Computational Biology and Chemistry* 78 (2019), 504–509.
[31] Meng-Yun Wu, Dao-Qing Dai, Xiao-Fei Zhang, and Yuan Zhu. 2013. Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm. *PloS one* 8, 6 (2013), e66256.
[32] Benhuai Xie, Wei Pan, and Xiaotong Shen. 2010. Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics* 26, 4 (2010), 501–508.
[33] Cong-Zhe You, Xiao-Jun Wu, Vasile Palade, and Abdulrahman Altahhan. 2016. Manifold locality constrained low-rank representation and its applications. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 3264–3271.