Published in final edited form as:

Genet Epidemiol. 2018 December ; 42(8): 796–811. doi:10.1002/gepi.22164.

Overlapping clustering of gene expression data using PWNCut

Sebastian J Teran Hidalgo¹, Tingyu Zhu², Mengyun Wu^{3,1,*}, and Shuangge Ma^{1,*}

¹Department of Biostatistics, Yale University

²Department of Statistics, Xiamen University

³School of Statistics and Management, Shanghai University of Finance and Economics

Abstract

Clustering has been widely conducted in the analysis of gene expression data. For complex diseases, it has played an important role in identifying unknown functions of genes, serving as the basis of other analysis, and others. A common limitation of most existing clustering approaches is to assume that genes are separated into disjoint clusters. As genes often have multiple functions and thus can belong to more than one functional cluster, the disjoint clustering results can be unsatisfactory. In addition, due to the small sample sizes of genetic profiling studies and other factors, there may not be sufficient evidence to confirm the specific functions of some genes and cluster them definitively into disjoint clusters. In this study, we develop an effective overlapping clustering approach, which takes account into the multiplicity of gene functions and lack of certainty in practical analysis. A penalized weighted normalized cut (PWNCut) criterion is proposed based on the NCut technique and an L_2 norm constraint. It outperforms multiple competitors in simulation. The analysis of TCGA data on breast cancer and cervical cancer leads to biologically sensible findings which differ from those using the alternatives. To facilitate implementation, we develop the function *pwncut* in the R package *NCutYX*.

Keywords

Gene expression data; Overlapping clustering; Penalization; NCut

1 Introduction

In biomedical studies, a huge amount of gene expression data have been collected and successfully used for understanding the etiology and progression of many complex diseases. Clustering is an important step in gene expression data analysis. It can identify biologically relevant gene clusters, which has significant implications in the discovery of unknown gene functions, comprehension of biological/cellular processes, and others (Basford, McLachlan, & Rathnayake, 2012). For example, the Glucocorticoid-Induced TNF receptor has been

*Correspondence: Mengyun Wu, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China. wu.mengyun@mail.shufe.edu.cn. Shuangge Ma, Department of Biostatistics, Yale University, New Haven, CT, USA shuangge.ma@yale.edu.

Conflict of Interest

The authors declare no conflict of interest.

demonstrated to play a functional role in regulating the CD4⁺CD25⁺ T cell subset by using the self-organizing map clustering algorithm (McHugh et al., 2002). Clustering can also serve as the basis of other analysis, such as reducing dimensionality in regression models (Li, Cook, & Nachtsheim, 2004; Bartenhagen, Klein, Ruckert, Jiang, & Dugas, 2010).

Many clustering approaches have been proposed in the literature and applied to gene expression data, including hierarchical clustering, K-means, graph-based clustering, model-based clustering, and others (Andreopoulos, An, Wang, & Schroeder, 2009; Xu & Wunsch, 2010; Wiwie, Baumbach, & Röttger, 2015). Among them, graph-based clustering has received much attention in recent years (Yu, Wong, & Wang, 2007; Lee, Zaïane, Park, Huang, & Greiner, 2008; Yu et al., 2016). Such analysis is often based on a two-stage process, where a graph is first constructed based on a similarity matrix and a graph-based partitioning criterion is then developed for the final clustering. Different approaches may have differences in terms of similarity definitions, partitioning criteria and optimization algorithms, but share the same spirit of maximizing similarity within clusters while minimizing similarity across clusters. Benefiting from the definitions and concepts in graph theory, such as the spectral graph theory, graph-based clustering approaches often have a solid mathematical ground and effective computational algorithms. Some approaches also have the advantage of preserving the local structure of the nodes from a high dimensional space. We refer to Schaeffer (2007), Nascimento & De Carvalho (2011), and others for discussions.

Despite considerable successes, many of the existing clustering approaches may still be limited in assuming disjoint clusters. It is commonly recognized that many genes have several biological functions and thus belong to more than one functional cluster. Take BRCA1, a well-known breast cancer gene which is also analyzed in the BRCA data in this article, as an example. It has been suggested to be involved in seven KEGG pathways, including the PI3K-Akt signaling pathway, Breast cancer, Fanconi anemia pathway, and others. Recently, multiple families of approaches have been developed to tackle the problem of overlapping clusters (N'Cir, Cleuziou, & Essoussi, 2015; Baadel, Thabtah, & Lu, 2016). In the literature, a classic strategy is to represent the cluster memberships using weights in the [0,1] interval, which describe the belonging probability of each gene to each cluster. With this strategy, the most representative approach is perhaps fuzzy c-means (FCM) (Bezdek, Ehrlich, & Full, 1984; Dembele & Kastner, 2003), under which a weighted inertia criterion is proposed. Other relevant developments include the fuzzy clustering by local approximation of membership (FLAME) (Fu & Medico, 2007), semi-supervised fuzzy clustering algorithm (SSFCA) (Maraziotis, 2012), GO fuzzy relational clustering (GO-FRC) (Paul & Shill, 2018), and others. These fuzzy clustering approaches can facilitate the identification of overlapping clusters with easy interpretations. In addition, they are usually based on partitioning strategies with simple similarity measures, which have lower computational complexity (compared to some alternatives, for example, hierarchical clustering) and are suitable for high-dimensional gene expression data (Baadel, Thabtah, & Lu, 2016). In practical studies, there is sometimes not sufficient evidence to definitively identify the specific functions of genes because of small sample sizes, high noises, and other reasons. A conservative strategy, which does not make a firm statement on cluster membership unless there is overwhelming evidence, can be desirable. However, the existing

approaches, including the fuzzy clustering described above, usually do not pay enough attention to this problem and thus may generate overly “bold” clustering results. In addition, it has been demonstrated in “classic” low dimensional data analysis that there is no dominating clustering approach (Baadel, Thabtah, & Lu, 2016). All the aforementioned factors call for the development of new and more effective clustering approaches.

In this study, our goal is to develop a more effective clustering approach for gene expression data. Different from K-means, hierarchical clustering and other disjoint clustering analysis, the proposed approach allows for overlapping clusters and can be more flexible. Significantly different from the existing overlapping clustering analysis, special attention is paid to the “lack of sufficient information” resulted from the limited sample size and other factors. Different from FCM-related approaches, a novel penalized weighted normalized cut (NCut) criterion is proposed, built on the NCut technique and an L_2 norm similarity constraint. NCut is one of the most popular graph-based clustering approaches and has a solid statistical ground and satisfactory numerical performance. To facilitate practical application, we develop the function *pwncut* in the R package *NCutYX*. With both methodological and numerical advancements, the proposed approach can provide a practically powerful new venue for clustering analysis.

2 Methods

Consider a dataset with n independent subjects. Denote $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})$ as the p –dimensional vector of gene expression (GE) measurements for subject i and $\mathbf{X}$ as the $n \times p$ design matrix composed of $\mathbf{X}_i$ ’s.

2.1 Clustering analysis with NCut

The proposed approach is based on the NCut technique. For the completeness of this article and to facilitate presentation, we first briefly describe the NCut technique and refer to the literature (Shi & Malik, 2000) for more details. A $p \times p$ similarity matrix $\mathbf{W} = (w_{jl})_{p \times p}$ is first constructed for the p GEs, with the non-negative element w_{jl} measuring the similarity between genes j and l . In this study, the absolute Pearson correlation $|\text{cor}(\mathbf{X}_j, \mathbf{X}_l)|$ is adopted for defining w_{jl} due to its simplicity and intuitive interpretation, where $\mathbf{X}_j$ is the length- n vector composed of the GE measurements for the n subjects and gene j . Other measures, such as the absolute distance correlation and inverse of Euclidean distance, can also be applied. Denote $\mathcal{A}_1, \dots, \mathcal{A}_K$ as the index sets of K disjoint clusters. The ordinary NCut measure is defined as

$$\text{NCut}(\mathcal{A}_1, \dots, \mathcal{A}_K) = \sum_{k=1}^K \frac{\sum_{j \in \mathcal{A}_k, l \in \mathcal{A}_k^c} w_{jl}}{\sum_{j, l \in \mathcal{A}_k} w_{jl}}, \quad (1)$$

where $\mathcal{A}_k^c$ is the complement of $\mathcal{A}_k$. With a fixed K , the optimal clustering minimizes the NCut measure.

NCut is first developed for image segmentation (Shi & Malik, 2000) and more recently has demonstrated promising performance in the analysis of gene expression data (Xing & Karp, 2001; Chen & Jian, 2014). It has a very intuitive formulation that minimizes the across-cluster similarity (numerator) and maximizes the within-cluster similarity (denominator) simultaneously. Another advantage is that no assumptions are made on the similarity measure or underlying data distribution and models. However, like many other existing clustering approaches, NCut has the limitation of assuming disjoint clusters.

2.2 Clustering analysis with penalized weighted NCut

Different from the traditional NCut approach described above, a *weight* matrix $A = (A_1, \dots, A_K)_{p \times K}$ is introduced in the proposed approach, where $A_k = (a_{1k}, \dots, a_{pk})'$ and $0 \leq a_{jk} \leq 1$ measures the probability of the j th gene belonging to cluster k with $\sum_{k=1}^K a_{jk} = 1$. This weighted strategy shares a similar spirit with that of fuzzy clustering. We propose the PWNCut (Penalized Weighted NCut) objective function as

$$\text{PWNCut}(A; W, \lambda) = \sum_{k=1}^K \frac{\text{wcut}(A_k, 1 - A_k; W)}{\text{wvol}(A_k; W)} + \lambda \sum_{j=1}^p \sum_{k=2}^K (a_{j,(k)} - a_{j,(k-1)})^2, \quad (2)$$

where

$$\text{wcut}(A_k, 1 - A_k; W) = A_k' W (1 - A_k), \quad (3)$$

$$\text{wvol}(A_k; W) = A_k' W A_k, \quad (4)$$

$\mathbf{1} = (1, 1, \dots, 1)'$, $a_{j(1)} \ a_{j(2)} \ \dots \ a_{j(K-1)} \ a_{j(K)}$ are the ordered elements of $A_j = (a_{j1}, \dots, a_{jK})$, and $\lambda \geq 0$ is a data-dependent tuning parameter. With fixed K and λ , the proposed estimate of A minimizes the objective function (2), subject to the constraints

$$0 \leq a_{jk} \leq 1 \text{ and } \sum_{k=1}^K a_{jk} = 1. \quad (5)$$

Here, when the estimated weights $a_{j1}, \dots, a_{jK}$ are close to each other, there is not sufficient evidence to determine the unique cluster membership for gene j ; on the other hand, when one weight is large enough and far away from others, there is dominating evidence to assign the corresponding cluster label to gene j . If specific cluster memberships are desired, a simple thresholding can be applied to A , where subjects having weights that exceed this threshold are classified into the corresponding specific clusters.

Rationale The proposed approach has been motivated by the following considerations. The first term in (2) is a flexible “upgrade” of the NCut technique. If the constraints in (5) are replaced by

$$a_{jk} \in \{0, 1\} \text{ and } \sum_{k=1}^K a_{jk} = 1,$$

the first term goes back to the ordinary NCut measure (1). Compared to many existing clustering approaches that assign only one cluster label to each gene, the proposed approach allows genes to belong to more than one cluster using a weight matrix, and can lead to more flexible clustering results. The proposed weighted strategy is partly motivated by fuzzy clustering. Significantly different from the existing fuzzy clustering analysis, the second term is innovatively proposed to shrink weights towards equal. This strategy is specially designed to accommodate the “lack of sufficient information” resulted from small sample size and other reasons. The proposed approach takes a conservative strategy in the sense that it does not assign a specific label to a gene unless there is overwhelming evidence. The penalization strategy has also been adopted in the clustering literature, for example the sparsity constrained k -extended clustering (Lu, Hong, Street, Wang, & Tong, 2012). It is based on the matrix decomposition technique where an L_1 penalty is adopted to make the weight matrix $\mathbf{A}$ sparser. Note that with unique considerations, the proposed similarity penalty differs significantly from the existing studies. Specifically, the proposed approach may better accommodate the multiplicity of gene functions and lack of certainty in practical analysis with small/moderate sample sizes. Note that in the current context with the special form of the penalty (on adjacent weights), it is not desirable to generate sparse estimates. As such, the L_1 and other sparse penalties are not needed. The tuning parameter λ is introduced in a similar spirit as other penalization approaches. Specifically, a larger λ makes a_{jk} 's ($k = 1, \dots, K$) closer to each other. In the extreme case where $\lambda \rightarrow \infty$, all weights are shrunk to $1/K$ so that each gene belongs to all clusters with equal probabilities. On the other hand, a smaller λ makes the estimated weights more dependent on the NCut term and more distant from each other.

2.3 Computation

With fixed K and λ , we adopt the simulated annealing (SA) technique (Bertsimas & Tsitsiklis, 1993) for optimizing the objective function (2). SA is an iterative algorithm and has been a popular choice in many published studies. The proposed algorithm proceeds as follows.

Step 1 Initialize $t = 0$ and $A^{(t)} = (a_{jk}^{(t)})_{p \times K}$ with all entries equal to $1/K$, where $A^{(t)}$ denotes the estimate of $\mathbf{A}$ at iteration t .

Step 2 Set $t = t + 1$. Draw $k(-)$ and $k(+)$ from $\{1, \dots, K\}$ randomly with probability $1/K$ and j from $\{1, \dots, p\}$ randomly with probability $1/p$. Generate u from the Uniform distribution $\text{Unif}(0, a_{j, k(-)}^{(t-1)})$. Set

$$a_{j,k(-)}^{(t)} = a_{j,k(-)}^{(t-1)} - u \text{ and } a_{j,k(+)}^{(t)} = a_{j,k(+)}^{(t-1)} + u.$$

For $k' = k(+), k(-)$ and $j \neq j'$, set $a_{jk'}^{(t)} = a_{jk'}^{(t-1)}$.

Step 3 If $\text{PWNCut}(A^{(t)}; W, \lambda) \leq \text{PWNCut}(A^{(t-1)}; W, \lambda)$, keep $A^{(t)}$ as it is. If not, keep

$A^{(t)}$ as it is with probability $\exp\left(-\frac{\text{PWNCut}(A^{(t)}; W, \lambda) - \text{PWNCut}(A^{(t-1)}; W, \lambda)}{T(t)}\right)$, and

otherwise $A^{(t)} = A^{(t-1)}$, where $T(t) = L \log(t+1)$ is the temperature function at iteration t with initial temperature L . Following published literature (Yip & Pao, 2009), we set $L = 100$.

Step 4 Repeat Steps 2–3 until $t = t_{\max}$, where $t_{\max}$ is the maximum number of iterations.

Step 5 Return the estimated matrix $A^{(t_{\max})}$ at iteration $t_{\max}$.

The SA technique has been used successfully along with the NCut-based approaches and shown to have satisfactory performance (Hidalgo, Wu, & Ma, 2017; Hidalgo & Ma, 2018). In Step 2, a possible solution is generated randomly and evaluated by the comparison with that in the previous iteration. The proposed generating approach for u and update strategy for a_{kj} ensure that the constraints in (5) are satisfied. Convergence properties of the SA technique with a large enough value of $t_{\max}$ have been intensively studied in the literature (Bertsimas & Tsitsiklis, 1993). In our numerical studies, we set $t_{\max} = 30,000$, which generates satisfactory results.

The proposed algorithm has complexity $O(npKt_{\max})$ and is computationally affordable. For example, with fixed K , ten values of λ and $t_{\max} = 30,000$, for a simulated dataset with $p = 500$ and $n = 300$, the proposed analysis takes about 2 minutes on a laptop with standard configurations. **Tuning parameter selection** There are two tuning parameters in the proposed objective function. Besides the number of clusters K which is commonly involved in clustering analysis, an additional parameter λ is introduced. We select K using the GAP statistic (Tibshirani, Walther, & Hastie, 2001), which has been widely adopted in clustering analysis and can be implemented using the R function *clusGap*. For choosing λ , we adopt V -fold cross-validation. To simplify description, take $V = 2$ as an example. We randomly split the n subjects into two sets $\mathcal{N}_1$ and $\mathcal{N}_2$ with equal sizes. For each λ , the proposed approach is trained on $\mathcal{N}_1$ and tested on $\mathcal{N}_2$, followed by being trained on $\mathcal{N}_2$ and tested on $\mathcal{N}_1$. We can next obtain the estimated weight matrices $A(\mathcal{N}_1, \lambda)$ and $A(\mathcal{N}_2, \lambda)$. Then, the cross-validation error is defined as

$$CV(\lambda) = \frac{1}{np} \left(\left\| X_{\mathcal{N}_2} - X_{\mathcal{N}_2} P(\mathcal{N}_1, \lambda) \right\|_F^2 + \left\| X_{\mathcal{N}_1} - X_{\mathcal{N}_1} P(\mathcal{N}_2, \lambda) \right\|_F^2 \right), \quad (6)$$

Where $\|\cdot\|_F$ is the Frobenious norm, $X_{\mathcal{N}}$ denotes the rows of X indexed by set $\mathcal{N}$, and

$$P(\mathcal{N}_l, \lambda) = A(\mathcal{N}_l, \lambda) [A'(\mathcal{N}_l, \lambda) A(\mathcal{N}_l, \lambda)]^{-1} A'(\mathcal{N}_l, \lambda) \quad (7)$$

is the projection matrix for the l th ($l = 1, 2$) training run. The optimal value of λ is chosen by minimizing the cross-validation error $CV(\lambda)$. This measure shares a similar spirit with the K-means objective function developed in Boutsidis, Zouzias, & Drineas (2010) for dimension reduction, which is based on the random projection approach. The proposed measure (6) aims to minimize the distance of each subject from its corresponding weighted cluster center.

A Toy example We further consider a toy example with 100 subjects to investigate the operating characteristics of the proposed approach. There are 50 GEs forming two overlapping clusters, with 20 GEs belonging to both clusters. Data are simulated under Scenario S2 with $\alpha_1 = \alpha_2 = 1$, but with a lower dimensionality, where α_1 and α_2 describe the membership weights for genes belonging to two clusters (see the next section for details). Under this setting, the 20 GEs (that belong to two clusters) have equal membership weights ($a_{j1} = a_{j2}$). The true values of a_{j1} 's, together with the estimated values using the proposed approach, K-means clustering (KM) and fuzzy c-means clustering (FCM), are shown in Figure 1. The two alternatives are perhaps the most representative approaches for disjoint and overlapping clustering. The values of a_{j1} 's are indicated by different colors. It is observed that the proposed approach can more accurately identify the underlying cluster memberships. Specifically, it separates the two clusters clearly and finds the overlaps correctly. Compared to the proposed approach, both KM and FCM fail to discover the genes belonging to both clusters. More conclusive results are presented in the next section.

2.4 NCutYX R package

We have developed an R function *pwncut* for implementing the proposed clustering approach, which is available as part of the R package *NCutYX* (<https://cran.r-project.org/web/packages/NCutYX/index.html>). Specifically, the proposed function *pwncut* can be implemented as

```
clust ← pwncut(X, K = 4, B = 1000, L = 1000, dist = 'correlation', lambda = 1).
```

This function has six inputs, where “X” is the design matrix for GEs, “K” is the number of clusters, “B” is the maximum number of SA iterations, “L” is the initial temperature, “dist” specifies that the absolute Pearson correlation is used for similarity, and “lambda” is the tuning parameter. The resulting object “clust” is a list where the first entry “clust[[1]]” is a vector of SA sequence and the second entry “clust[[2]]” consists of the estimated weight matrix A . The grid search for the tuning parameters can be easily realized based on *pwncut*.

3 Simulation

Performance of the proposed approach is evaluated using extensive simulations. The design matrix $\mathbf{X}$ is simulated as follows. Given the number of clusters K , we first generate vector $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})'$, $i = 1, \dots, n$, from the multivariate Normal distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{2I})$ with mean $\boldsymbol{\mu} = (0, \dots, 0)'$ and covariance $\mathbf{I}_{K \times K}$, where $\mathbf{I}_{K \times K}$ is the $K \times K$ identity matrix. Then, given the coefficient matrix $\mathbf{C} = (c_{jk})_{p \times K}$, let

$$X_{i.} = (\mathbf{C}\mathbf{z}_i)' + \varepsilon_i, \quad i = 1, \dots, n,$$

where each element ε_{ij} ($i = 1, \dots, n, j = 1, \dots, p$) of $\boldsymbol{\varepsilon}_i$ is iid from the Normal distribution $\mathcal{N}(0, 1)$. The true weight matrix $\mathbf{A}^*$ is defined as the normalization of matrix $\mathbf{C}$ with $\sum_{k=1}^K a_{jk}^* = 1$. The following six specific scenarios are considered.

S1 has 400 genes with four overlapping clusters and

$$\mathbf{C} = \begin{pmatrix} 1_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} \\ \alpha_1 1_{100 \times 1} & \alpha_2 1_{100 \times 1} & 0_{100 \times 1} & 0_{100 \times 1} \\ 0_{25 \times 1} & 1_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} \\ 0_{25 \times 1} & 0_{25 \times 1} & 1_{25 \times 1} & 0_{25 \times 1} \\ 0_{100 \times 1} & 0_{100 \times 1} & \alpha_1 1_{100 \times 1} & \alpha_2 1_{100 \times 1} \\ 0_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} & 1_{25 \times 1} \end{pmatrix}, \quad (8)$$

where α_1 and α_2 describe the membership weights for genes belonging to two clusters. For example, if $\alpha_1 = \alpha_2$, then the 26th to 125th genes have equal membership weights for clusters 1 and 2. There are 100 genes belonging to both clusters 1 and 2, and another 100 genes belonging to both clusters 3 and 4.

S2 has 500 genes with two overlapping clusters and

$$\mathbf{C} = \begin{pmatrix} 1_{200 \times 1} & 0_{200 \times 1} \\ \alpha_1 1_{100 \times 1} & \alpha_2 1_{100 \times 1} \\ 0_{200 \times 1} & 1_{200 \times 1} \end{pmatrix}. \quad (9)$$

There are 100 genes belonging to both clusters.

S3 has 300 genes with three overlapping clusters and

$$C = \begin{pmatrix} 1_{90 \times 1} & 0_{90 \times 1} & 0_{90 \times 1} \\ \alpha_1 1_{10 \times 1} & \alpha_2 1_{10 \times 1} & \alpha_3 1_{10 \times 1} \\ 0_{90 \times 1} & 1_{90 \times 1} & 0_{90 \times 1} \\ \alpha_1 1_{10 \times 1} & \alpha_2 1_{10 \times 1} & \alpha_3 1_{10 \times 1} \\ 0_{90 \times 1} & 0_{90 \times 1} & 1_{90 \times 1} \\ \alpha_1 1_{10 \times 1} & \alpha_2 1_{10 \times 1} & \alpha_3 1_{10 \times 1} \end{pmatrix}. \quad (10)$$

There are 30 genes belonging to all three clusters.

S4 has 120 genes with four overlapping clusters and

$$C = \begin{pmatrix} 1_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} \\ \alpha_1 1_{10 \times 1} & \alpha_2 1_{10 \times 1} & 0_{10 \times 1} & 0_{10 \times 1} \\ 0_{25 \times 1} & 1_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} \\ 0_{25 \times 1} & 0_{25 \times 1} & 1_{25 \times 1} & 0_{25 \times 1} \\ 0_{10 \times 1} & 0_{10 \times 1} & \alpha_1 1_{10 \times 1} & \alpha_2 1_{10 \times 1} \\ 0_{25 \times 1} & 0_{25 \times 1} & 0_{25 \times 1} & 1_{25 \times 1} \end{pmatrix}. \quad (11)$$

There are 10 genes belonging to both clusters 1 and 2, and another 10 genes belonging to both clusters 3 and 4.

S5 has 600 genes with 12 overlapping clusters. Here, for the first 100 genes and the first two clusters, the sub-matrix of C is equal to

$$\begin{pmatrix} 1_{40 \times 1} & 0_{40 \times 1} \\ \alpha_1 1_{20 \times 1} & \alpha_2 1_{20 \times 1} \\ 0_{40 \times 1} & 1_{40 \times 1} \end{pmatrix}. \quad (12)$$

The other five sub-matrices have the same patterns as the one defined above, and the remaining elements are zero. There are 20 genes belonging to both clusters k and $k + 1$ for $k = 1, \dots, 11$.

S6 has 120 genes with four disjoint clusters and

$$C = \begin{pmatrix} 1_{30 \times 1} & 0_{30 \times 1} & 0_{30 \times 1} & 0_{30 \times 1} \\ 0_{30 \times 1} & 1_{30 \times 1} & 0_{30 \times 1} & 0_{30 \times 1} \\ 0_{30 \times 1} & 0_{30 \times 1} & 1_{30 \times 1} & 0_{30 \times 1} \\ 0_{30 \times 1} & 0_{30 \times 1} & 0_{30 \times 1} & 1_{30 \times 1} \end{pmatrix}. \quad (13)$$

For each scenario, we examine multiple settings with various values of n and α_k . A total of 44 simulation settings are considered, comprehensively covering a wide spectrum with various numbers of genes, subjects and clusters, and different patterns of overlapping (overlapping or disjoint, different membership weights, etc.).

3.1 Parameter path

Since the proposed penalization strategy differs significantly from those in the existing clustering studies, we first examine the parameter path (as a function of λ) to better comprehend the effects of penalization. We simulate one replicate with 200 subjects under each of the four different simulation settings, including Scenario S2 with $\alpha_1 = \alpha_2 = 1$, Scenario S3 with $\alpha_1 = \alpha_2 = \alpha_3 = 1$, Scenario S2 with $\alpha_1 = 3$ and $\alpha_2 = 1$, and Scenario S3 with $\alpha_1 = \alpha_2 = 3$ and $\alpha_3 = 1$. Under the former two settings, genes belonging to multiple clusters have equal membership weights, whereas under the latter two settings, those genes have unequal membership weights. Scenarios S2 and S3 have two and three overlapping clusters, respectively.

The parameter paths are shown in Figure A.1. For each gene, the estimated weights $a_{j1}, \dots, a_{jK}$ are represented by lines with the same color. The vertical lines correspond to the values of λ chosen by cross-validation. It is observed that the membership weights for each gene are shrunk towards $1/K$ when λ increases, as expected by design. In addition, the proposed cross-validation can lead to estimation with satisfactory properties. For example, for Scenario S1 with $\alpha_1 = \alpha_2 = 1$ (top left panel of Figure A.1), two types of genes are observed with the cross-validation selected λ . The first one (blue) has weights equal to zero and one. The second type (pink) has both weights approximately equal to 0.5. This result is in accordance with the setting in (9). With the cross-validation selected λ , two types of genes are also found for Scenario S1 with $\alpha_1 = 3$ and $\alpha_2 = 1$ (bottom left panel of Figure A.1). Different from the former one, genes belonging to two clusters have weights approximately equal to 0.75 and 0.25, respectively, which is consistent with the true underlying structures. Similar patterns are observed for the other two datasets with $K = 3$.

3.2 Comparison with the alternative approaches

Beyond the proposed approach, the following alternatives are also considered. (a) K-means clustering (**KM**), which partitions p genes into K disjoint clusters by minimizing the pairwise distances of genes in the same clusters. (b) Spectral clustering (**SC**), which uses the spectrum (eigenvalues) of the similarity matrix to obtain disjoint clusters (Ng, Jordan, & Weiss, 2002). (c) Fuzzy c-means clustering (**FCM**), which is an extension of the c-means clustering and generates fuzzy partitions and overlapping clusters (Bezdek, Ehrlich, & Full, 1984). It can be realized using the R package *e1071*. (d) Fuzzy K-means with entropy regularization (**FKME**), which is an overlapping clustering based on the maximum-entropy principle (Li & Mukaidono, 1995) and can be realized using the R package *fclust*. KM and SC are the most popular clustering approaches. Comparing with these two can reveal the value of the weighted strategy for overlapping clustering. FCM and FKME are two well-known clustering approaches that also allow for overlapping clusters. Comparing with these two can reveal the value of the proposed penalization strategy. We acknowledge that many other clustering approaches can also be adopted for analyzing the simulated data. We choose

the above alternatives due to their popularity and satisfactory performance, as well as their similar frameworks with the proposed. The comparison can in a relatively direct way establish the merit of the proposed weighted and penalization analysis.

To evaluate clustering performance, we use the following measure based on the estimated weight matrix A and true weight matrix A^* ,

$$M(A, A^*) = \sum_{j,l=1}^p |d_{jl} - d_{jl}^*|/p^2, \quad (14)$$

where

$$D = \sum_{k=1}^K A_k A_k' - \text{diag} \left(\sum_{k=1}^K A_k A_k' \right) + I_{p \times p} \triangleq (d_{jl})_{p \times p}$$

and

$$D^* = \sum_{k=1}^K A_k^* (A_k^*)' - \text{diag} \left(\sum_{k=1}^K A_k^* (A_k^*)' \right) + I_{p \times p} \triangleq (d_{jl}^*)_{p \times p},$$

with $\text{diag}(\cdot)$ containing the diagonal elements of the corresponding matrix. $M(A, A^*)$ measures the deviation between A and A^* , with a smaller value indicating better clustering performance.

For each setting, the mean and standard deviation of $M(A, A^*)$ over 1000 replicates are computed. The results for Scenarios S1 and S2 are presented in Tables 1 and 2, respectively, and the results for the others are presented in Appendix. Across the whole range of simulation settings, the proposed approach is observed to have superior or comparable performance. Under Scenario S1, a half of the GEs belong to multiple clusters, which may favor the proposed approach. The significant advantage of the proposed approach is observed with various values of n and α_1 . For example, with $n = 600$ and $\alpha_1 = \alpha_2 = 1$ (genes belonging to multiple clusters have equal membership weights), the proposed approach has mean $M(A, A^*) = 14.5$, compared to 29.1 (KM), 28.8 (SC), 24.0 (FCM), and 28.9 (FKME). It behaves better than KM and SC, which provides a strong support to the proposed weighted strategy. Compared to the two overlapping clustering approaches FCM and FKME, the proposed also has better performance, suggesting the effectiveness of the penalization strategy. Different from Scenario S1, the other scenarios have fewer genes belonging to multiple clusters. However, the proposed approach still achieves satisfactory results. Specifically, the advantage of the proposed approach is also prominent under Scenario S2. For example, with $n = 600$ and $\alpha_1 = \alpha_2 = 1$, the proposed approach has mean $M(A, A^*) = 1.2$, compared to 29.3 (KM), 31.6 (SC), 7.0 (FCM), and 17.9 (FKME). For Scenario S2, compared to the setting with $\alpha_1 = \alpha_2 = 1$, the proposed approach has a larger $M(A, A^*)$ under the setting with $\alpha_1 = 3$ and $\alpha_2 = 1$ (genes belonging to multiple clusters have unequal membership weights). This is as expected since the proposed approach encourages equal

weights, and $a_{j1} = a_{j2} = 0.5$ when $\alpha_1 = \alpha_2 = 1$. For Scenarios S3-S5 with a larger number of clusters, similar overall patterns are observed with the proposed approach having competitive performance. For Scenario S6, the proposed approach behaves slightly worse than the two disjoint clustering approaches KM and SC. This is reasonable as the true cluster structure is disjoint. However, the proposed approach still outperforms the other two overlapping clustering approaches FCM and FKME. For example, with $n = 100$, the proposed approach has mean $M(\mathbf{A}, \mathbf{A}^*) = 8.2$, compared to 3.9 (KM), 3.1 (SC), 24.9 (FCM), and 15.1 (FKME). We have also examined some other scenarios, and the observed patterns are similar (details omitted).

4 Data analysis

We analyze TCGA (<https://cancergenome.nih.gov/>) data on breast invasive carcinoma (BRCA) and cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC). TCGA data have been recently collected and published with a high quality, thus can provide an ideal testbed. We analyze the processed level 3 mRNA gene expression data, which are collected using the IlluminaHiSeq RNAseq V2 platform and have been lowess-normalized, log-transformed, and median-centered. The corresponding Z-scores represent the gene expression status (up or down regulated) in tumor samples relative to normal. The datasets have been downloaded from TCGA Provisional using the R package *cgdsr*.

4.1 BRCA Data

Although applicable in principle, the proposed analysis may not be reliable when the number of GEs is very large and sample size is limited. To improve stability, we conduct prescreening as follows. In TCGA, measurements are also available for a limited number of proteins. GEs that correspond to those selected “interesting” proteins can be potentially “more interesting”. As such, we select the top 1,000 GEs that have the strongest distance correlations with the protein measurements (Szekely & Rizzo, 2013). The distance correlation is an appropriate measure, as it can accommodate complex associations between GEs and proteins and has been demonstrated to have satisfactory performance in the literature (Hidalgo & Ma, 2018). Then, we use GOTerm Finder to identify 334 genes with well-defined GO terms for downstream analysis (Gene Ontology Consortium, 2014). This step of screening is conducted to increase interpretability. Data are available on 873 subjects.

With the GAP statistic, the proposed approach identifies eight clusters. The sums of weights $\sum_{j=1}^p a_{jk}$ of these clusters are 14.0, 46.2, 64.2, 47.5, 27.0, 28.0, 45.0 and 62.1. We show the estimated weight matrix $\mathbf{A} = (a_{jk})_{p \times K}$ in Figure 2. The values of a_{jk} 's are represented with different colors as indicated by the colorbar. It is observed that many genes belong to multiple clusters, which is justified by the multiple functionalities of most of the genes and establishes the necessity of the proposed overlapping strategy. Seven out of the eight clusters have overlapping genes.

To see whether the clusters are biologically meaningful, we further examine their functionalities based on the GO Panther “biological process” term (Gene Ontology Consortium, 2014). We select the top 20 processes with p-values less than 0.01 and separate

them into five categories: positive regulation metabolic (PRM) process, negative regulation metabolic (NRM) and biosynthetic (NRB) processes, metabolic (M) process (without a well-defined “direction”), and others (O). We calculate the weighted proportion of GEs in each cluster that have the corresponding process, and show the results in Figure 3, where the proportion is weighted by the estimated a_{jk} . With the proposed approach, differences across the eight clusters are clearly observed. More specifically, we list the enriched processes for each cluster in Table 3, demonstrating that different clusters represent different biological processes. For example, clusters 1, 6 and 8 have higher percentages of the NRB, NRM and PRM processes, respectively. In addition, in Figure 4, we display the weights of the eight genes that have the largest $\sum_{k=2}^K (a_{j,(k)} - a_{j,(k-1)})^2$. These genes have more prominently different weight distributions. We conduct analysis on their GO functions to examine the biological sense of overlapping clusters. For example, gene COA3 has been found to be over-represented in the positive regulation of cellular metabolic, macromolecule metabolic, metabolic and nitrogen compound metabolic processes. These processes are also over-represented in clusters 4, 7 and 8, which is consistent with the clustering results that COA3 has nonzero weights for clusters 4, 7 and 8. Similar biologically sensible observations are also made for many other genes, which provides support to the validity of the proposed approach.

Beyond the proposed approach, we also analyze data using the alternatives. To make different approaches more comparable, we fix the number of clusters as eight. The similarity of the analysis results is investigated based on the following concordance measure. Given two estimated weight matrices $\mathbf{A} = (a_{jk})_{p \times K}$ and $\mathbf{B} = (b_{jk})_{p \times K}$, we define

$$H(\mathbf{B}, \mathbf{A}) = \frac{\sum_{j=1}^p \sum_{k=1}^K a_{jk} b_{jk}}{\sum_{j=1}^p \sum_{k=1}^K a_{jk}}, \quad (15)$$

where a larger $H(\mathbf{B}, \mathbf{A})$ suggests a higher similarity. Note that this concordance measure is not symmetric. The summary comparison results are provided in Table A.5. It is observed that the proposed approach identifies cluster memberships different from those using the alternatives, and different approaches have moderate concordance. Results of FKME are not shown as it does not converge for this specific dataset.

4.2 CESC Data

A similar prescreening as described in the previous section is conducted, and 325 genes are selected for downstream analysis. Data are available on 164 subjects. With the GAP statistic, the proposed approach identifies four overlapping clusters, of which the sums of weights are 73.4, 89.7, 80.7 and 81.0. The heatmap for the estimated weight matrix is provided in Figure A.2. All four clusters have overlapping genes. We also analyze the GO biological processes for these genes, among which 17 processes have p-values less than 0.01. The weighted proportions of GEs in the four clusters that have the 17 processes are presented in Figure A.3, and the enriched processes for the four clusters are provided in Table A.6. It is observed that different clusters contain genes representing different processes. For example, cluster 1

has a higher percentage of adhesion or cellular processes, and cluster 2 has a higher percentage of reproduction related processes. The weights of the eight genes that have the largest $\sum_{k=2}^K (a_{j,(k)} - a_{j,(k-1)})^2$ are shown in Figure A.4. Different weight distributions are observed for these genes. For example, it is found that genes COX6A2 and MIR607 have nonzero weights for clusters 1, 3 and 4, while genes ANKRD49 and MIR5692B have nonzero weights for clusters 1, 2 and 4. Analysis is also conducted using the alternatives. The summary comparison results are shown in Table A.7. Different approaches are observed to identify different cluster memberships, and the levels of overlapping information as measured by $H(\mathbf{B}, \mathbf{A})$ are moderate.

5 Discussion

Clustering of gene expression data has been widely conducted and led to many promising findings for complex diseases (Beer et al., 2002; Calon et al., 2015). The existing approaches often do not take sufficient account of the unique characteristics of gene expression data, including the multiplicity of gene expressions and lack of sufficient information resulted from small sample size and other factors in practical data analysis. Thus, the clustering results so generated are still often unsatisfactory. In this study, we have developed a penalized weighted NCut approach for overlapping clustering analysis. The proposed approach introduces a cluster membership weight vector for each gene and takes advantage of the NCut technique to accommodate overlapping clusters. The most significant advancement from the existing approaches is that special attention is paid to the “lack of sufficient information” problem. Specifically, a novel penalization approach is developed to encourage “similarity” in weights. The proposed approach takes a conservative strategy, where genes that do not have enough evidence to be classified into specific clusters tend to have similar membership weights. An efficient SA algorithm has been developed to optimize the proposed objective function. Under a wide spectrum of simulation settings, the proposed approach is observed to have superior or comparable clustering performance compared to the alternative disjoint and overlapping clustering approaches. In the analysis of two TCGA datasets, it identifies clusters different from the alternatives. The GO term analysis shows significant differences across the biological functions of different clusters. These findings provide an indirect support to the validity of the proposed clustering.

This study can be potentially extended in multiple directions. Although the NCut technique has multiple advantages, it can be of interest to develop clustering based on other techniques, for example spectral clustering or K-means. The proposed weighted and penalization strategy can be potentially coupled with other clustering techniques. With the proposed penalty, a gene is encouraged to belong to multiple clusters unless there is strong evidence. Different penalties can be developed to address other prior considerations, and it is conjectured that a similar SA algorithm can be proposed. The SA technique has been commonly adopted in the existing penalization studies with various penalties, including the L_1 penalty (Gramacy & Polson, 2012), smoothly clipped absolute deviation (SCAD) penalty (Fan & Peng, 2004), and others. Other stochastic optimization techniques, such as the genetic algorithm and tabu search, may also be applicable to the proposed approach (Fouskakis & Draper, 2002). They have been widely conducted in the clustering literature

and shown to have good performance, in particular in the accommodation of large-scale gene expression data (Bandyopadhyay, Mukhopadhyay, & Maulik, 2007; Rahman & Islam, 2014; Sung & Jin, 2000). In cross-validation for choosing λ , other measures can also be adopted besides the proposed error (6). For example, it may be possible to maximize the Davies-Bouldin-index (DBI) (Davies & Bouldin, 1979), which is the ratio of the across-cluster distances to within-cluster distances and has been a popular choice in clustering analysis (Gupta, Mehrotra, & Mohan, 2010; Pacella, Grieco, & Blaco, 2016). The proposed approach can also be potentially extended to analyze multiple datasets from independent studies with comparable designs. Following the literature (Liu, Huang, & Ma, 2013; Huang, Liu, Yi, Shia, & Ma, 2017), effects of the same gene from multiple datasets can be treated as a group, and then a group-based penalty can be introduced to accommodate the heterogeneity among multiple datasets. This extension is expected to be nontrivial and may demand a separate development. Besides gene expression data, the proposed overlapping clustering is also potentially applicable to other types of omics data as well as data in other fields, such as somatic mutation data, image data and text data. Minor modifications on the definition of similarity matrix may be needed to accommodate specific properties of other data types. For example, entropy can be adopted for discrete somatic mutation data. Extensive simulations and data analysis show that the proposed approach has satisfactory performance. More extensive numerical analysis as well as confirmation studies of the data analysis findings may be of interest.

Acknowledgments

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of the article. This work was partly supported by the National Institutes of Health [CA216017, CA204120]; National Natural Science Foundation of China [61402276, 91546202]; and Bureau of Statistics of China [2016LD01].

Appendix

Figure A.1:

Parameter path for (a) Scenario S2 with $\alpha_1 = \alpha_2 = 1$, (b) Scenario S2 with $\alpha_1 = 3$ and $\alpha_2 = 1$, (c) Scenario S3 with $\alpha_1 = \alpha_2 = \alpha_3 = 1$, (d) Scenario S3 with $\alpha_1 = \alpha_2 = 3$ and $\alpha_3 = 1$. The vertical lines correspond to the values of λ chosen by cross-validation.

Figure A.2:

Analysis of CESC data: heatmap of the estimated weight matrix $A = (a_{jk})_{p \times K}$. The values of a_{jk} 's are represented with different colors as indicated by the colorbar.

Figure A.3:

Analysis of CESC data: the weighted proportions of GEs with a certain GO process in the four clusters. The GO processes are abbreviated for a better display.

Figure A.4:
 Analysis of CESC data: weights of the eight genes that have the largest $\sum_{k=2}^K (a_{j,(k)} - a_{j,(k-1)})^2$.

Table A.1:

Simulation Scenario S3. In each cell, mean $M(A, A^*)$ (sd) over 1000 replicates.

<i>n</i>	<i>a</i> ₁	<i>a</i> ₂	<i>a</i> ₃	PWNCut	KM	SC	FCM	FKME
30	1	1	1	9.7(4.5)	15.2(9.4)	16.3(9.3)	9.0(1.6)	8.7(1.0)
100	1	1	1	8.0(0.9)	16.5(10.7)	15.5(9.6)	8.3(0.6)	8.7(0.8)
300	1	1	1	7.8(0.6)	15.2(8.9)	14.6(9.0)	8.2(0.3)	8.6(0.7)
600	1	1	1	7.7(0.6)	15.3(9.0)	14.0(9.0)	8.2(0.2)	8.6(0.1)
30	3	2	1	9.7(3.0)	18.8(11.6)	21.0(7.6)	12.6(8.4)	24.6(1.3)
100	3	2	1	8.8(0.6)	18.3(11.9)	20.1(8.1)	11.7(8.2)	24.5(0.9)
300	3	2	1	8.6(0.7)	18.8(12.2)	20.0(8.9)	11.2(7.9)	24.5(0.1)
600	3	2	1	8.5(0.6)	19.3(12.8)	20.8(8.5)	11.3(8.0)	24.5(0.1)

Table A.2:

Simulation Scenario S4. In each cell, mean $M(A, A^*)$ (sd) over 1000 replicates.

<i>n</i>	<i>a</i> ₁	<i>a</i> ₂	PWNCut	KM	SC	FCM	FKME
30	1	1	4.0(3.1)	12.6(5.9)	12.2(5.7)	7.8(1.6)	7.5(0.4)

n	a_1	a_2	PWNCut	KM	SC	FCM	FKME
100	1	1	3.1(0.6)	12.3(5.1)	11.7(5.6)	7.8(1.5)	18.3(0.1)
300	1	1	3.1(0.5)	12.7(5.1)	11.5(5.1)	7.8(1.4)	18.3(0.1)
600	1	1	3.0(0.4)	12.8(5.7)	11.4(5.0)	7.7(1.4)	18.3(0.1)
30	3	1	9.2(1.3)	28.4(6.0)	24.9(9.1)	11.9(7.7)	14.4(0.1)
100	3	1	12.1(2.2)	29.8(4.4)	22.8(9.1)	11.5(7.9)	16.7(2.4)
300	3	1	12.0(2.2)	30.1(4.3)	21.6(9.6)	11.3(8.1)	16.0(0.1)
600	3	1	12.0(2.1)	29.9(4.3)	21.3(8.9)	10.8(7.8)	16.0(0.1)

Table A.3:

Simulation Scenario S5. In each cell, mean $M(A, A^*)$ (sd) over 1000 replicates.

n	a_1	a_2	PWNCut	KM	SC	FCM	FKME
30	1	1	4.3(0.6)	5.8(1.6)	6.2(2.2)	4.9(3.1)	3.6(0.6)
100	1	1	4.2(1.0)	5.5(1.4)	5.2(1.6)	5.5(3.2)	4.7(0.3)
300	1	1	3.6(0.6)	5.4(1.4)	4.9(1.2)	5.4(2.9)	4.1(0.1)
600	1	1	2.6(0.9)	5.5(1.3)	4.7(0.9)	5.1(2.5)	3.8(0.1)
30	3	1	8.7(0.5)	9.8(2.4)	11.2(4.0)	6.6(6.6)	10.8(1.2)
100	3	1	8.6(0.5)	10.0(2.9)	9.7(4.2)	8.7(5.7)	9.8(0.9)
300	3	1	8.4(0.4)	10.1(2.8)	8.9(5.0)	9.3(5.8)	9.3(0.2)
600	3	1	8.2(0.5)	9.6(2.4)	7.2(2.7)	9.4(6.2)	9.2(0.2)

Table A.4:

Simulation Scenario S6. In each cell, mean $M(A, A^*)$ (sd) over 1000 replicates.

n	PWNCut	KM	SC	FCM	FKME
30	7.9(1.5)	3.5(6.7)	3.1(6.1)	16.7(2.8)	16.1(3.1)
100	8.2(1.9)	3.9(6.7)	3.1(6.3)	24.9(0.8)	15.1(2.4)
300	9.0(1.4)	4.0(7.3)	2.7(6.1)	17.6(0.3)	15.4(0.1)
600	7.7(1.3)	3.9(7.0)	2.9(6.2)	11.3(0.7)	15.5(0.1)

Table A.5:

Analysis of BRCA data: comparison of the clustering results by different approaches. In each cell, $H(B, A)$, where B and A are the weight matrices estimated using the approaches in the column and row, respectively.

	PWNCut	KM	SC	FCM
PWNCut	100%	57.6%	86.0%	55.9%
KM	16.4%	100%	87.4%	85.8%
SC	16.2%	57.9%	100%	54.7%
FCM	17.2%	93.3%	89.9%	100%

Table A.6:

Analysis of CESC data: the enriched processes for the four clusters.

Process	Cluster			
	1	2	3	4
Biological adhesion				
Catabolic process				
Cell-cell adhesion				
Cell adhesion				
Cellular process				
Chromatin organization				
Chromosome segregation				
Fertilization				
Gamete generation				
Immune system process				
Nitrogen compound metabolic process				
Reproduction				
Response to toxic substance				
Sensory perception				
Sensory perception of chemical stimulus				
Spermatogenesis				
Steroid metabolic process				

Table A.7:

Analysis of CESC data: comparison of the clustering results by different approaches. In each cell, $H(\mathbf{B}, \mathbf{A})$, where $\mathbf{B}$ and $\mathbf{A}$ are the weight matrices estimated using the approaches in the column and row, respectively.

	PWNCut	KM	SC	FCM	FKME
PWNCut	100%	46.5%	73.3%	71.2%	43.0%
KM	33.6%	100%	91.9%	73.1%	61.3%
SC	27.8%	48.3%	100%	68.0%	35.7 %
FCM	33.3%	47.3%	83.9%	100%	39.7 %
FKME	40.6%	80.2%	88.8%	79.9%	100%

References

- Andreopoulos B, An A, Wang X, & Schroeder M (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics*, 10, 297–314. *doi:10.1093/bib/bbn058* [PubMed: 19240124]
- Baadel S, Thabtah F, & Lu J (2016, 7). Overlapping clustering: A review. In *SAI Computing Conference (SAI)*, 2016 (pp. 233–237). IEEE. *doi:10.1109/SAI.2016.7555988*
- Bandyopadhyay S, Mukhopadhyay A, & Maulik U (2007). An improved algorithm for clustering gene expression data. *Bioinformatics*, 23(21), 2859–2865. *doi:10.1093/bioinformatics/btm418* [PubMed: 17720981]

- Bartenhagen C, Klein HU, Ruckert C, Jiang X, & Dugas M (2010). Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics*, 11, 567. doi:10.1186/1471-2105-11-567 [PubMed: 21087509]
- Basford KE, McLachlan GJ, & Rathnayake SI (2012). On the classification of microarray gene-expression data. *Briefings in Bioinformatics*, 14, 402–410. doi:10.1093/bib/bbs056 [PubMed: 22988257]
- Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, ... & Lizyness ML (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8), 816–824. doi:10.1038/nm733
- Bertsimas D, & Tsitsiklis J (1993). Simulated annealing. *Statistical Science*, 8(1), 10–15. doi: 10.1214/ss/1177011077
- Bezdek JC, Ehrlich R, & Full W (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10, 191–203. doi: 10.1016/0098-3004(84)90020-7
- Boutsidis C, Zouzias A, & Drineas P (2010). Random projections for *k*-means clustering. In *Advances in Neural Information Processing Systems* (pp. 298–306).
- Calon A, Lonardo E, Berenguer-Llargo A, Espinet E, Hernando-Momblona X, Iglesias M, ... & Cortina C (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature Genetics*, 47(4), 320–329. doi:10.1038/ng.3225 [PubMed: 25706628]
- Chen X, & Jian C (2014). Gene expression data clustering based on graph regularized subspace segmentation. *Neurocomputing*, 143, 44–50. doi: 10.1016/j.neucom.2014.06.023
- Davies DL, & Bouldin DW (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2, 224–227. doi:10.1109/TPAMI.1979.4766909
- Dembele D, & Kastner P (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics*, 19, 973–980. doi: 10.1093/bioinformatics/btg119 [PubMed: 12761060]
- Fan J, & Peng H (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3), 928–961. doi:10.1214/009053604000000256
- Fouskakis D, & Draper D (2002). Stochastic optimization: a review. *International Statistical Review*, 70(3), 315–349. doi:10.1111/j.1751-5823.2002.tb00174.x
- Fu L, & Medico E (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, 8, 3. doi: 10.1186/1471-2105-8-3 [PubMed: 17204155]
- Gene Ontology Consortium. (2014). Gene ontology consortium: going forward. *Nucleic Acids Research*, 43, D1049–D1056. doi:10.1093/nar/gku1179 [PubMed: 25428369]
- Gramacy RB, & Polson NG (2012). Simulation-based regularized logistic regression. *Bayesian Analysis*, 7(3), 567–590. doi:10.1214/12-BA719
- Gupta A, Mehrotra KG, & Mohan C (2010). A clustering-based discretization for supervised learning. *Statistics & Probability Letters*, 80(9–10), 816–824. doi:10.1016/j.spl.2010.01.015
- Hidalgo SJT, Wu M, & Ma S (2017). Assisted clustering of gene expression data using ANCut. *BMC Genomics*, 18, 623. doi:10.1186/s12864-017-3990-1 [PubMed: 28814280]
- Hidalgo SJT, & Ma S (2018). Clustering multilayer omics data using MuNCut. *BMC Genomics*, 19, 198. doi:10.1186/s12864-018-4580-6 [PubMed: 29703159]
- Huang Y, Liu J, Yi H, Shia BC, & Ma S (2017). Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. *Statistics in Medicine*, 36(3), 509–559. doi:10.1002/sim.7138 [PubMed: 27667129]
- Lee CH, Zaïane OR, Park HH, Huang J, & Greiner R (2008). Clustering high dimensional data: a graph-based relaxed optimization approach. *Information Sciences*, 178, 4501–4511. doi:10.1016/j.ins.2008.05.014
- Li RP, & Mukaidono M (1995, 3). A maximum-entropy approach to fuzzy clustering. In *Fuzzy Systems, 1995*. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int (Vol. 4, pp. 2227–2232). IEEE. doi:10.1109/FUZZY.1995.409989
- Li L, Cook RD, & Nachtshiem CJ (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics & Data Analysis*, 47(1), 175–193. doi:10.1016/j.csda.2003.10.017

- Liu J, Huang J, & Ma S (2013). Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statistics in Medicine*, 32(20), 3509–3521. *doi:10.1002/sim.5780* [PubMed: 23519988]
- Lu H, Hong Y, Street WN, Wang F, & Tong H (2012, 12). Overlapping clustering with sparseness constraints. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* (pp. 486–494). IEEE. *doi:10.1109/ICDMW.2012.16*
- Maraziotis IA (2012). A semi-supervised fuzzy clustering algorithm applied to gene expression data. *Pattern Recognition*, 45, 637–648. *doi:10.1016/j.patcog.2011.05.007*
- McHugh RS, Whitters MJ, Piccirillo CA, Young DA, Shevach EM, Collins M, & Byrne MC (2002). CD4+ CD25+ immunoregulatory T cells: gene expression analysis reveals a functional role for the glucocorticoid-induced TNF receptor. *Immunity*, 16, 311–323. *doi:10.1016/S1074-7613(02)00280-7* [PubMed: 11869690]
- Nascimento MC, & De Carvalho AC (2011). Spectral methods for graph clustering-a survey. *European Journal of Operational Research*, 211, 221–231. *doi:10.1016/j.ejor.2010.08.012*
- NCir CEB, Cleuziou G, & Essoussi N (2015). Overview of overlapping partitional clustering methods In *Partitional Clustering Algorithms* (pp. 245–275). Springer, Cham.
- Ng AY, Jordan MI, & Weiss Y (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (pp. 849–856).
- Pacella M, Grieco A, & Blaco M (2016). On the use of self-organizing map for text clustering in engineering change process analysis: a case study. *Computational Intelligence and Neuroscience*, 2016, 1–11. *doi:10.1155/2016/5139574*
- Paul AK, & Shill PC (2018). Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. *Biosystems*, 163, 1–10. *doi:10.1016/j.biosystems.2017.09.017* [PubMed: 29113811]
- Rahman MA, & Islam MZ (2014). A hybrid clustering technique combining a novel genetic algorithm with K-Means. *Knowledge-Based Systems*, 71, 345–365. *doi:10.1016/j.knosys.2014.08.011*
- Schaeffer SE (2007). Graph clustering. *Computer Science Review*, 1(1), 27–64. *doi:10.1016/j.cosrev.2007.05.001*
- Shi J, & Malik J (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 888–905. *doi:10.1109/34.868688*
- Sung CS, & Jin HW (2000). A tabu-search-based heuristic for clustering. *Pattern Recognition*, 33(5), 849–858. *doi:10.1016/S0031-3203(99)00090-4*
- Szkely GJ, & Rizzo ML (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117, 193–213. *doi:10.1016/j.jmva.2013.02.012*
- Tibshirani R, Walther G, & Hastie T (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411–423. *doi:10.1111/1467-9868.00293*
- Wiwie C, Baumbach J, & Röttger R (2015). Comparing the performance of biomedical clustering methods. *Nature Methods*, 12, 1033–1038. [PubMed: 26389570]
- Xing EP, & Karp RM (2001). CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, 17, S306–S315. *doi:10.1093/bioinformatics/17.suppl1.S306* [PubMed: 11473022]
- Xu R, & Wunsch DC (2010). Clustering algorithms in biomedical research: a review. *IEEE Reviews in Biomedical Engineering*, 3, 120–154. *doi:10.1109/RBME.2010.2083647* [PubMed: 22275205]
- Yang MS, & Nataliani Y (2017). Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. *Pattern Recognition*, 71, 45–59. *doi:10.1016/j.patcog.2017.05.017*
- Yip PP, & Pao YH (1995). Combinatorial optimization with use of guided evolutionary simulated annealing. *IEEE Transactions on Neural Networks*, 6(2), 290–295. *doi:10.1109/72.363466* [PubMed: 18263313]
- Yu Z, Wong HS, & Wang H (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, 23, 2888–2896. *doi:10.1093/bioinformatics/btm463* [PubMed: 17872912]

Yu Z, Luo P, You J, Wong HS, Leung H, Wu S, ... & Han G (2016). Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 28, 701–714. *doi:10.1109/TKDE.2015.2499200*

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

: Toy example: genes are marked with different colors based on the values of a_{jl} 's, where for example, red, purple and blue represent 0, 0.5 and 1, respectively. The distance between two genes is computed based on Pearson correlation.

Figure 2.

: Analysis of BRCA data: heatmap of the estimated weight matrix $A = (a_{jk})_{p \times K}$. The values of a_{jk} 's are represented with different colors as indicated by the colorbar.

Figure 3.
: Analysis of BRCA data: the weighted proportions of GEs with a certain GO process in the eight clusters. The GO processes are abbreviated for a better display.

Figure 4.

: Analysis of BRCA data: weights of the eight genes that have the largest

$$\sum_{k=2}^K \left(a_{j,(k)} - a_{j,(k-1)} \right)^2.$$

Table 1

: Simulation Scenario S1. In each cell, mean $M(A, A^*)$ (sd) over 1000 replicates.

n	α_1	α_2	PWNCut	KM	SC	FCM	FKME
30	1	1	14.9(0.5)	29.4(4.1)	29.5(4.6)	24.1(1.7)	29.6(0.7)
100	1	1	14.7(0.3)	29.3(4.4)	28.9(4.3)	24.0(1.6)	28.9(0.1)
300	1	1	14.6(0.2)	29.3(4.3)	28.7(4.6)	24.0(1.5)	28.9(0.1)
600	1	1	14.5(0.2)	29.1(4.3)	28.8(4.4)	24.0(1.4)	28.9(0.1)
30	3	1	18.9(1.9)	28.5(5.4)	28.1(4.9)	25.7(0.2)	22.3(2.0)
100	3	1	17.9(0.1)	28.6(6.0)	27.9(5.5)	25.6(0.2)	26.6(1.0)
300	3	1	17.8(0.2)	28.6(6.1)	27.9(5.2)	25.5(1.3)	26.5(1.3)
600	3	1	17.8(0.2)	28.9(6.6)	27.9(5.6)	25.4(0.1)	26.1(0.1)

Table 2

: Simulation Scenario S2. In each cell, mean $M(\mathbf{A}, \mathbf{A}^*)$ (sd) over 1000 replicates.

n	α_1	α_2	PWNCut	KM	SC	FCM	FKME
30	1	1	9.2(3.4)	28.7(7.5)	30.3(7.0)	9.7(0.1)	31.9(0.1)
100	1	1	1.7(0.1)	28.0(7.7)	30.4(7.4)	7.8(0.2)	10.0(0.1)
300	1	1	1.3(0.1)	28.6(7.5)	29.4(9.0)	7.1(0.1)	16.9(0.1)
600	1	1	1.2(0.1)	29.3(7.3)	31.6(6.8)	7.0(0.1)	17.9(0.1)
30	3	1	9.3(0.6)	28.3(7.1)	28.8(7.8)	42.4(15.3)	15.4(0.1)
100	3	1	9.4(0.1)	28.8(7.1)	29.4(7.4)	47.0(9.6)	17.9(0.1)
300	3	1	9.4(0.1)	26.1(7.5)	28.2(6.7)	49.2(3.1)	17.9(0.1)
600	3	1	9.4(0.1)	25.5(7.6)	27.4(7.7)	49.4(0.1)	17.9(0.1)

Table 3

: Analysis of BRCA data: the enriched processes for the eight clusters.

Process	Cluster							
	1	2	3	4	5	6	7	8
Cell proliferation								
Metabolic								
Mono car boxy lie acid metabolic								
NR biosynthetic								
NR cellular biosynthetic								
NR cellular macromolecule biosynthetic								
NR cellular metabolic								
NR macromolecule metabolic								
NR metabolic								
NR nitrogen compound metabolic								
Organic acid metabolic								
Organic substance metabolic								
Oxoacid metobolic								
PR cellular metabolic								
PR macromolecule metabolic								
PR metabolic								
PR nitrogen compound metabolic								
PR nucleobase compound metabolic								
Small molecule catabolic								
Small molecule metabolic								