Journal of Data Science 20 (1), 14–33
January 2022

DOI: 10.6339/21-JDS1028
Statistical Data Science

Integrative Clustering Analysis with Application in Multi-Source Gene Expression Data

Liuqing Yang[1], Qing Pan[1], and Yunpeng Zhao[2,*]

[1]*Department of Statistics, George Washington University, Washington, D.C., U.S.A.*
[2]*School of Mathematical and Natural Sciences, Arizona State University, Tempe, Arizona, U.S.A.*

Abstract

In omics studies, different sources of information about the same set of genes are often available. When the group structure (e.g., gene pathways) within the genes are of interests, we combine the normal hierarchical model with the stochastic block model, through an integrative clustering framework, to model gene expression and gene networks jointly. The integrative framework provides higher accuracy in extensive simulation studies when one or both of the data sources contain noises or when different data sources provide complementary information. An empirical guideline in the choice between integrative versus separate clustering models is proposed. The integrative clustering method is illustrated on the mouse embryo single cell RNAseq and bulk cell microarray data, which identified not only the gene sets shared by both data sources but also the gene sets unique in one data source.

Keywords *EM algorithm; empirical guidelines; microarray data; normal hierarchical model; single cell RNAseq; stochastic block model*

1 Introduction

Network analysis is the study of networks representing relationships (i.e., links or edges) between objects (i.e., vertices or nodes). Examples of networks include social networks, World Wide Web, protein-protein interactions, logistics supply chains, etc. A large number of probabilistic and statistical models have been proposed (Goldenberg et al., 2010). In this paper, we focus on community structure in networks, where a community is a set of nodes that is densely connected internally and sparsely connected to the rest of the network.

Community structures are often reflected in the gene expression levels when genes in the same cluster are expressed synergistically. Normal hierarchical model (NHM) is a special case of Bayesian hierarchical models where individual observations follow normal distributions whose mean parameters share a common prior (Morris and Lysy, 2012). We modify the classical NHM by assuming one NMH in each gene cluster. That is, mean expression levels from genes in different clusters follow normal prior distributions with different parameters.

The normal hierarchical model (NHM) is often employed in analyzing microarray data (Thompson et al., 2020). However, single cell RNAseq data cannot be easily modeled by standard probability distributions due to high levels of noises, dropouts, outliers and over-dispersion. Therefore, we dichotomize connectivity measures in single cell RNAseq data and apply stochastic block model (SBM) on the resulting gene network data. SBM is a popular community detection approach on binary network data, which assumes that the link probability of each pair of nodes

*Corresponding author. Email: yunpeng.zhao@asu.edu.

Received April 6, 2021; Accepted October 12, 2021

is determined by their community labels (Holland et al., 1983). A lot of progresses have been made on the theoretical justification and computational methods of SBM (Bickel and Chen, 2009; Karrer and Newman, 2011; Zhao et al., 2012; Bickel and Chen, 2009; Amini et al., 2013; Abbe, 2017; Zhao, 2017). In gene network data, a pair of genes are considered to be connected if their expressions are tightly related. The similarity of two gene expression profiles may be measured by the Euclidean distance (Priness et al., 2007) or correlation coefficients (Zhang and Horvath, 2005). The main goal is to partition the gene network into cohesive groups, which may coincide with functional gene sets or pathways. Binary links in a gene network are defined by setting thresholds on the distances or correlation values. The estimated group structure is usually sensitive to the choice of the threshold and would be less reliable if an inappropriate threshold was chosen (Perkins and Langston, 2009).

With more and more microarray and single cell RNA (scRNA) data in the public data consortium, integrative analysis combining these two data sources has achieved higher accuracy and power in differential expression analysis (Forcato et al., 2021). Different sources of gene expression data on the same set of genes collected under similar conditions often provide complementary information on the underlying structure of gene sets or pathways. But there lacks an integrative clustering method explicitly designed for joint analysis of microarray and scRNA data. In multi-omics studies, integrative clustering methods can be roughly classified into five categories (Rappoport and Shamir, 2018): concatenating matrices (Wu et al., 2015; Wang et al., 2013), clustering omics separately followed by integration of clusters (Hoadley et al., 2014; Nguyen et al., 2017), integrating similarities (Wang et al., 2014), dimension reduction followed by clustering (Lock et al., 2013; Zhang et al., 2012) and probability clustering (Mo et al., 2013; Lock and Dunson, 2013). The proposed method falls into the category of probability clustering models, since it combines the log likelihoods of the NHM and the SBM on two sets of data from independent data sources. Specifically, the NHM describes the clustering structure on the mean values of gene expression levels, while the SBM extracts groups using mutual distances between genes. The integrative log-likelihood equals the sum of the log-likelihood from a hierarchical model on the microarray data and the log-likelihood from a stochastic block model on the binary network from scRNA data.

The rest of the paper is organized as follows. In Section 2, a novel EM algorithm is proposed, which combines the pseudo-likelihood method (Amini et al., 2013) for the SBM and the NHM. Extensive simulation studies are carried out in Section 3. We examine the performance of integrative clustering versus separate clustering in the presence of contamination as well as orthogonal community structures in different data sources. In most scenarios the proposed integrative method has a higher accuracy in identifying the latent community structure compared to the methods using a single data source. Furthermore, we provide empirical guidelines for the choice between integrative analysis and separate analysis. In Section 4, our integrative clustering model is applied to a microarray data and an scRNA data, measuring gene expression levels in mouse embryo cells under the same experimental conditions, independently generated in two different labs. Section 5 discusses limitations and future research directions.

2 Methods

2.1 Normal Hierarchical Model for Gene Expression Data

We adopt NHM to model the normalized counts of gene expression. Specifically, the means of gene-specific expression levels share group-level prior distributions. We first set up the notations.

Suppose there are p genes and n subjects. The expression level of gene i ($i = 1, \ldots, p$) in subject j ($j = 1, \ldots, n$) (after proper transformations) is denoted by x_{ij}, and let $X = [x_{ij}]$. Let $c = (c_1, \ldots, c_p)$ where c_i is the community label of gene i and there are K non-overlapping communities in total. Let $P(c_i = k) = \pi_k, k = 1, \ldots, K, \pi = (\pi_1, \ldots, \pi_K)^T$. We further assume x_{ij} follows $N(\mu_{ik}, 1/\gamma_i)$ given $c_i = k$, where the gene-specific precision γ_i is treated as a fixed parameter, which is estimated using observation from gene i. And μ_{ik} follows a prior distribution $N(\mu_k, 1/\eta_k)$, where μ_k and η_k are the group-level mean and precision, respectively. Let $\theta_k = (\mu_k, \eta_k)$. The likelihood of the NHM is

$$f(x_{ij}|c_i = k, \gamma_i, \mu_{ik}) = \frac{1}{\sqrt{2\pi}}\gamma_i^{\frac{1}{2}} \exp\left(-\frac{(x_{ij} - \mu_{ik})^2\gamma_i}{2}\right),$$
$$f(\mu_{ik}|\theta_k) = \frac{1}{\sqrt{2\pi}}\eta_k^{\frac{1}{2}} \exp\left(-\frac{(\mu_{ik} - \mu_k)^2\eta_k}{2}\right).$$

Furthermore, let $x_i = (x_{i1}, \ldots, x_{in})$, and the probability density function for the posterior distribution of μ_{ik} given x_i is

$$\begin{aligned} f(\mu_{ik}|x_i, c_i = k, \theta_k, \gamma_i) &\propto f(x_i|c_i = k, \mu_{ik}, \gamma_i) f(\mu_{ik}|\mu_k, \eta_k) \\ &= (2\pi)^{-\frac{n}{2}}(\gamma_i)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\gamma_i \sum_{j=1}^{n}(x_{ij} - \mu_{ik})^2\right\} \times \\ &\quad (2\pi)^{-\frac{1}{2}}(\eta_k)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\eta_k(\mu_{ik} - \mu_k)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\eta_k + n\gamma_i)\left(\mu_{ik} - \frac{\eta_k\mu_k + \gamma_i \sum x_{ij}}{\eta_k + n\gamma_i}\right)^2\right\}. \end{aligned}$$

Since the prior on μ_{ik} is a conjugate prior, the posterior distribution is still normal. That is,

$$\mu_{ik}|x_i, c_i = k, \theta_k, \gamma_i \sim N\left(\frac{\eta_k\mu_k + \gamma_i \sum x_{ij}}{\eta_k + n\gamma_i}, \frac{1}{\eta_k + n\gamma_i}\right).$$

Finally, the probability density function of x_i conditional on $c_i = k$ (with μ_{ik} being integrated out), denoted as $f_X(x_i|\theta_k, c_i = k)$, $k = 1, \ldots, K$, goes as follows

$$\begin{aligned} f_X(x_i|\theta_k, c_i = k) = (2\pi)^{-\frac{n}{2}} &\frac{\gamma_i^{\frac{n}{2}}\eta_k^{\frac{1}{2}}}{(\eta_k + n\gamma_i)^{\frac{1}{2}}} \times \\ &\exp\left\{-\frac{1}{2}\frac{\gamma_i}{\eta_k + n\gamma_i}\left[\eta_k \sum_{j=1}^{n}(x_{ij} - \mu_k)^2 + n\gamma_i \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2\right]\right\}. \end{aligned} \quad (1)$$

We use the sample variance to estimate γ_i, that is, $[\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2]^{-1}$, and replace γ_i by the estimator in formulas hereafter. Next we apply the EM algorithm to obtain the MLE of $\{\pi_k\}$ and $\{\theta_k\}$.

Let

$$T_{k,i} = P(c_i = k|x_i, \{\pi_k\}, \{\theta_k\}) = \frac{f_X(x_i|\theta_k, c_i = k)\pi_k}{\sum_{l=1}^{K} f_X(x_i|\theta_l, c_i = l)\pi_l},$$
$$L(\{\pi_k\}, \{\theta_k\}; X, \{c_i\})] = f_X(X|\{\theta_k\}, \{c_i\})P(\{c_i\}|\{\pi_k\}), \quad k = 1, \ldots, K, \ i = 1, \ldots, p.$$

Then the expectation of the log-likelihood function in the t-th iteration of the EM algorithm, with respect to the conditional distribution of $\{c_i\}$ given X, $\{\pi_k\}$ and $\{\theta_k\}$ is

$$E_{\{c_i\}|X,\{\pi_k\}^{(t)},\{\theta_k\}^{(t)}}[\log L(\{\pi_k\}^{(t)},\{\theta_k\}^{(t)};X,\{c_i\})] = \sum_{i=1}^{p}\sum_{k=1}^{K} T_{k,i}^{(t)}[\log \pi_k^{(t)} + \log f(x_i|\theta_k^{(t)}, c_i = k)].$$

In the M-step,

$$\pi^{(t+1)} = \arg\max_{\pi} \sum_{i=1}^{p}\sum_{k=1}^{K} T_{k,i}^{(t)} \log \pi_k,$$

which implies

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{p} T_{k,i}^{(t)}}{\sum_{l=1}^{K}\sum_{i=1}^{p} T_{l,i}^{(t)}}, \quad k = 1,\ldots,K.$$

And

$$\begin{aligned}\theta_k^{(t+1)} &= \arg\max_{\theta_k=(\mu_k,\eta_k)} \sum_{i=1}^{p} T_{k,i}^{(t)} \log f_X(x_i|\theta_k, c_i = k) \\ &= \arg\max_{\theta_k} \sum_{i=1}^{p} T_{k,i}^{(t)} \left[\frac{1}{2}\log\left(\frac{\eta_k}{\eta_k + n\gamma_i}\right) + \frac{n}{2}\log\gamma_i - \frac{1}{2}\frac{\gamma_i\eta_k}{\eta_k+n\gamma_i}\sum_{j=1}^{n}(x_{ij}-\mu_k)^2 - \frac{1}{2}\frac{\gamma_i n(n-1)}{\eta_k+n\gamma_i}\right].\end{aligned}$$

Because the optimization problem above does not have a closed-form solution, BFGS quasi-Newton method (Fletcher, 2013) is employed using the R function optim.

2.2 Pseudo-Likelihood Method for the Stochastic Block Model

We adopt the stochastic block model (SBM) to model the community structure within a network with binary edges, that is, an adjacency matrix $A = [A_{ii'}]$ with $A_{ii'} = 1$ if node i and i' are connected, $A_{ii'} = 0$ otherwise. The SBM assumes that the linkage probability between a pair only depends on the community labels c_i and $c_{i'}$ of the two endpoints. It is well known that the E-step of a classical EM algorithm is intractable when applied to the SBM. Amini et al. (2013) proposed a scalable pseudo-likelihood method that can fit a large network under the SBM efficiently. The key idea is to approximate the original adjacency matrix by a sample from a mixture of multivariate Poisson distributions and fit the model by a standard EM algorithm. We summarize this method below. Let $e = (e_1,\ldots,e_p)$ be an initial community assignment. Define $b_{ik} = \sum_{i'} A_{ii'}1(e_{i'} = k)$ for $i = 1,\ldots,p$ and $k = 1,\ldots,K$. Let $b_i = (b_{i1},\ldots,b_{iK})$.

As observed by Amini et al. (2013), each b_i approximately follows a mixture of multivariate Poisson distribution. That is, given $c_i = k$, b_{il} is approximately Poisson with mean denoted by λ_{kl}, $l = 1,\cdots,K$. Moreover, as observed by Amini et al. (2013), $\{b_i\}$ are approximately independent across i because b_{ik} and $b_{i'l}$ $(i \neq i')$ share at most one common link, and are approximately independent when p is large. Letting $\lambda_k = (\lambda_{k1},\ldots,\lambda_{kK})$, the conditional probability of b_i given $c_i = k$ is (up to a constant)

$$f_A(b_i|\lambda_k, c_i = k) = \exp\left(-\sum_{l=1}^{K}\lambda_{kl}\prod_{l=1}^{K}\lambda_{kl}^{b_{il}}\right), \tag{2}$$

and the joint pseudo log-likelihood is (up to a constant)

$$\sum_{i=1}^{p} \log \left[\sum_{k=1}^{K} \pi_k \exp\left(-\sum_{l=1}^{K} \lambda_{kl} \right) \prod_{l=1}^{K} \lambda_{kl}^{b_{il}} \right].$$

A maximum pseudo-likelihood estimate of $\{\pi_k\}, \{\lambda_k\}$ can then be obtained via the standard EM algorithm for mixture models. Once the EM converges, the initial community assignment e is updated to the most current estimate of node labels. This process is repeated for a fixed number of iterations. In practice, Amini et al. (2013) observed that it usually takes a small number of updates to stabilize e.

Finally, we specify the choice of initial values $\{\pi_k^{(0)}\}, \{\lambda_k^{(0)}\}$ as proposed in Amini et al. (2013). Given the initial community assignment e obtained by spectral clustering,

$$\pi_k^{(0)} = \frac{1}{p} \sum_{i=1}^{p} 1(e_i = k), \tag{3}$$

$$\lambda_{kl}^{(0)} = \sum_{ii'} A_{ii'} 1(e_i = k, e_{i'} = l) / \sum_{i} 1(e_i = k). \tag{4}$$

2.3 Integrative Method

In the presence of both microarray data and the binary adjacent matrix, we combine the log-likelihood from the NHM and the log-pseudo-likelihood from the SBM to infer the underlying community structure. To specify the joint likelihood, we make the following two additional assumptions:

1. X and A share the same community labels c;
2. Given c, X and A are independent.

Recall that x_i is the row of X corresponding to gene i and b_i is a K-dimensional vector that measures the connections between node i and each block according to the initial labeling e. Based on the assumptions above, the integrative pseudo-likelihood for (x_i, b_i) given $c_i = k$ is

$$IG(x_i, b_i; \theta_k, \lambda_k) = f_X(x_i|\theta_k, c_i = k) f_A(b_i|\lambda_k, c_i = k),$$

where $f_X(x_i|\theta_k, c_i = k)$ and $f_A(b_i|\lambda_k, c_i = k)$ are defined in (1) and (2), respectively. As in Section 2.2, a maximum likelihood estimate of $(\{\pi_k\}, \{\theta_k\}, \{\lambda_k\})$ can be obtained via the EM algorithm. After the EM loop converges, we use the estimated posterior probabilities to update the labeling e and re-compute $\{b_i\}$.

We now give the detailed algorithm. Start with initial labeling e and initial parameters $(\{\pi_k^{(0)}\}, \{\theta_k^{(0)}\}, \{\lambda_k^{(0)}\})$. Then repeat the process below T times:

(1) Compute b_{ik} according to e as

$$b_{ik} = \sum_{i'} A_{ii'} 1(e_{i'} = k), \quad i = 1, \ldots, p, \quad k = 1, \ldots, K.$$

(2) Use current parameter estimates $\pi^{(t)}$, $\theta^{(t)}$ to compute the conditional probabilities for node labels as

$$T_{k,i}^{(t)} = P(c_i = k|x_i, b_i) = \frac{\pi_k^{(t)} IG(x_i, b_i; \theta_k^{(t)}, \lambda_k^{(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} IG(x_i, b_i; \theta_l^{(t)}, \lambda_l^{(t)})}, \quad i = 1, \ldots, p, \quad k = 1, \ldots, K.$$

(3) Given $T_{k,i}^{(t)}$, update $(\{\pi_k^{(t+1)}\}, \{\theta_k^{(t+1)}\}, \{\lambda_k^{(t+1)}\})$ as

$$\pi_k^{(t+1)} = \frac{1}{p}\sum_{i=1}^{p} T_{k,i}^{(t)}, \quad k = 1, \ldots, K,$$

$$\theta_k^{(t+1)} = \arg\max_{\theta_k} \sum_{i=1}^{p} T_{k,i}^{(t)} \log f_X(x_i|\theta_k, c_i = k), \quad k = 1, \ldots, K,$$

and

$$\lambda_{kl}^{(t+1)} = \frac{\sum_{i=1}^{p} T_{k,i}^{(t)} b_{il}}{\sum_{i=1}^{p} T_{k,i}^{(t)}}, \quad k = 1, \ldots, K, \quad l = 1, \ldots, K.$$

(4) Repeat steps (2) and (3) until convergence.
(5) Update labels by $e_i = \arg\max_k T_{k,i}^{(t+1)}$, $i = 1, \ldots, p$.

3 Simulation

In this section, we evaluated the performance of the proposed integrative method by simulation studies. We fixed $p = 1000$, $n = 40$, $K = 3$, $\pi = (0.3, 0.3, 0.4)$ and repeated $T = 12$ times in the outer loop of the integrative EM algorithm, which is sufficient for estimated labels reaching the stable status in most cases. Conditional on the labels, the gene expression data x_{ij} $(i = 1, \ldots, p,$ $j = 1, \ldots, n)$ were generated from the NHM, that is, $x_{ij}|c_i = k \sim N(\mu_{ik}, \gamma_i)$, where $\mu_{ik} \sim N(\mu_k, \eta_k)$. We chose $(\mu_1, \eta_1) = (-0.5, 1/0.36)$, $(\mu_2, \eta_2) = (2, 1/0.36)$, $(\mu_3, \eta_3) = (4, 1/0.64)$, and γ_i were independently generated by $1/(\text{Unif}(0.2, 1.5))^2$. Conditional on the labels, the edges of A were generated as independent Bernoulli variables with probability 0.12 if a pair of nodes are from the same community, otherwise with probability 0.06.

We further introduced contamination in X and A. Specifically, for X, let each $x_i = (x_{i1}, \ldots, x_{in})$ be replaced by an $n \times 1$ vector where each component was sampled from $N(2.5, 1)$ independently with probability $p.noise.X$. For A, we randomly chose pairs of nodes with probability $p.noise.A$, and set their linkage probability to be $p_0 = 0.0001$. Both $p.noise.X$ and $p.noise.A$ varied from 0.1 to 0.9 with increments 0.1.

For the NHM, we first applied k-means to find the initial labels and given the initial labels, we computed the maximum likelihood estimates and used them as $\{\theta_k^{(0)}\}$. For the SBM, the initial labels were produced by spectral clustering and $\{\pi_k^{(0)}\}$, $\{\lambda_k^{(0)}\}$ are computed according to (3) and (4). For the integrative model, we obtained initial values $\{\theta_k^{(0)}\}$ from X and $\{\pi_k^{(0)}\}$, $\{\lambda_k^{(0)}\}$ from A, respectively, by the aforementioned methods. The EM algorithm starts from an E-step using the initial parameter estimates.

To compare the estimated labels with the true labels, we use the adjusted Rand index (ARI), which is a measure of similarity between two partitions. Higher ARI indicates higher degree of agreement between two sets of labels (Rand, 1971). Each plot shows three ARIs, representing the performance of three methods: the NHM using X alone, the SBM using A alone, and the integrative method (IG) using both data types. Additionally, we plot the ARI measuring the similarity between labels estimated from NHM and SBM (denoted as "NHM vs SBM" in the legend).

Figure 1 shows the four sets of ARI values with varied $p.noise.X$ (from 0.1 to 0.9) and $p.noise.A$ (from 0.1 to 0.9). As expected, the performance of all three methods becomes worse as the contamination levels in two data sources increase. The integrative method performs

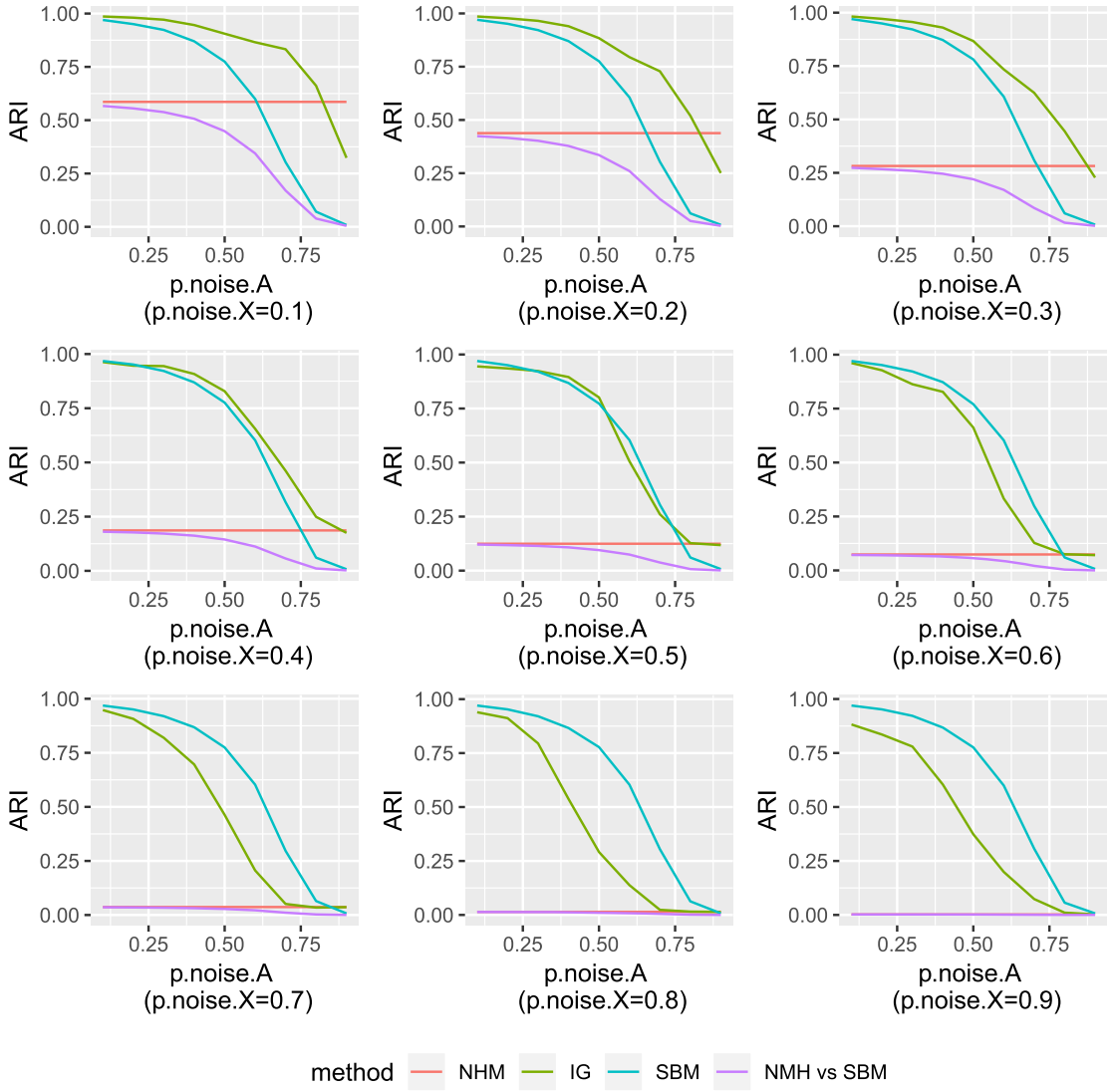

Figure 1: Comparison of the performance of integrative method vs SBM along vs NHM alone when X and A share a common community structure: the ARI between true/estimated and estimated labels as a function of $p.noise.X$ and $p.noise.A$.

better than both SBM and NHM when the contamination levels in X and A are from low to moderate, say both less than 0.4. On the other hand, when one data source contains a high level of contamination while the other has low contamination levels, for example, $p.noise.X = 0.1$ and $p.noise.A = 0.9$, the integrative model is no longer better than the method using the single source with little contamination. Without knowledge of which data source contains more contamination, the integrative method is a safe choice because it is always better than the worse one among the two data sources. However, if we know one data source is less contaminated than the other, we need to choose either the integrative or the better data source alone depending on the contamination levels. We therefore propose an empirical guideline in the choice between the integrative method and separate clustering analysis on individual data sources in the next section.

Table 1: Thresholds at different $p.noise.X$ levels, K from 3 to 8.

K	Threshold				
	$p.noise.X = 0.1$	$p.noise.X = 0.2$	$p.noise.X = 0.3$	$p.noise.X = 0.4$	Overall
3	0.039	0.026	0.016	0.180	0.180
4	0.059	0.049	0.213	0.147	0.213
5	0.024	0.021	0.014	0.084	0.084
6	0.032	0.024	0.018	0.006	0.032
7	0.020	0.016	0.011	0.007	0.020
8	0.015	0.009	0.007	0.005	0.015
3*	0.084	0.014	0.352	0.224	0.352
4*	0.058	0.046	0.031	0.163	0.163

3.1 Empirical Guidelines Choosing Between Integrative and Separate Analyses

Figure 1 indicates an interesting pattern across the four sets of ARIs. The ARI between the community estimates from A (SBM) and X (NHM) alone is small when the contamination level is heavy in either data source, which coincides with the performance pattern of the integrative model. That is, when the ARI between NHM and SBM is small, the integrative model is no longer the best. Therefore, we searched for a threshold in the ARI between NHM and SBM, below which the performance of our integrative model is inferior to at least one of the methods using a single data source. The first section of Table 1 lists the four thresholds from the first four subplots in Figure 1 when $K = 3$. As shown in the plots, for $p.noise.X = 0.1, 0.2, 0.3$, the integrative model performs the best in almost the entire range of $p.noise.A$, except for $p.noise.A = 0.9$. Thus the thresholds are equal to the second to the smallest ARI values between NHM and SBM at $p.noise.A = 0.9$. For $p.noise.X = 0.4$, the integrative model is the best when green line is above 0.18 at $p.noise.A = 0.1$, after which the integrative method is slightly worse than the SBM, and we employ the ARI between X and A at $p.noise.A = 0.1$ as the corresponding threshold. We do not show the thresholds for $p.noise.X \geqslant 0.5$ because our simulation (Figure 1) shows the integrative model is not the best even when $p.noise.A = 0.1$, and the ARIs between X and A are always lower than 0.12, which is the first ARI value of the green curve in Figure 1 at $p.noise.X = 0.5$. In summary, when $K = 3$ and the ARI between NHM and SBM is above 0.18, the integrative method using both data sources outperforms the clustering method using single source under our setup.

We carried out additional simulations to investigate how the threshold changes with the number of clusters K. We repeated the experiments as shown in Figure 1 under different K values, from 4 to 8. For each K, the true labels were evenly allocated, i.e., $\pi = (1/K, \ldots, 1/K)$. Parameters in the NHM, (μ_k, η_k) for $k = 1, 2, 3$ were the same as the settings in the previous simulation. For k from 4 to 8, we further set $\mu_4 = 1$, $\mu_5 = 3$, $\mu_6 = 5$, $\mu_7 = 6$, $\mu_8 = 7$ and $\eta_k = 1/0.36$. Moreover, the edges of A were generated as independent Bernoulli variables with probability 0.12 for within community pairs and probability 0.06 for between community pairs. The resulted thresholds versus the numbers of clusters are listed in blocks 2-6 of Table 1. The contaminations in A and X are generated in the same way as those in Figure 1. The column of overall thresholds lists the thresholds above which integrative analysis performs better for different K values. Table 1 shows that when K becomes larger, one can in general expect a

smaller threshold value. This may arise from the increasing possibilities of cluster memberships for each node when K is large, hence harder to reach an agreement between two sets of estimated clustering labels even though their data sources share a similar underlying clustering structure.

Lastly, the thresholds may depend on the parameter setting in the data generation process. Therefore, we recommend researchers run simulations based on the estimated values of parameters from their data to choose appropriate thresholds. We ran two sets of simulations, 3^* and 4^* in Table 1, to examine the robustness of thresholds from simulations using estimated parameters. We first generated X with $p.noise.X = 0.1$ and A with $p.noise.A = 0.1$ under the same setting as Section 3.1. We then fitted the NHM on X and the SBM on A respectively, and used the estimated parameters to generate data and find the thresholds for $K = 3$ and 4. The estimated parameters can be viewed as a perturbation of the original parameters. Although the thresholds in 3^* and 4^* are not identical to those in the two sets of simulations with $K = 3$ and $K = 4$, they are close numerically with the same decreasing trend.

3.2 Performance Under Unequal Community Structures

The previous simulation studies demonstrate that the integrative method pools different information sources and achieves higher accuracy when the ground truth of label c_X from X and c_A from A are identical, i.e., $c_{Xi} = c_{Ai}$ for $i = 1, \ldots, p$. Most of the existing integrative methods such as Wang et al. (2014), Yan and Sarkar (2021), Xu et al. (2012), and Newman and Clauset (2016) also assume the same underlying community structure among different data sources. However, under real life scenarios, whether different data sources share the same underlying community structure is unknown. Furthermore, it is possible that two different data sources contain different yet complementary community structures. In those cases, the overall community structure is the overlap of two sets of clusters. A pair of genes belongs to the same group in the overall community structure if and only if they belong to the same group in the clusters from X and also in the clusters from A. Below we demonstrate through simulations that the integrative method still provides higher accuracy in identifying the overall structure from two different sets of communities. Specifically, we investigate the performance of the integrative method given independent community labels c_X and c_A.

We first generated c_X and c_A independently with $K = 2$, both by p random draws from $Multinomial(1, (0.5, 0.5))$. Then we constructed the overall c with $K = 4$ where

$$c_i = \begin{cases} 1 & c_{Xi} = c_{Ai} = 1 \\ 2 & c_{Xi} = 1, c_{Ai} = 2 \\ 3 & c_{Xi} = 2, c_{Ai} = 1 \\ 4 & c_{Xi} = c_{Ai} = 2 \end{cases}, \quad i = 1, \ldots, p.$$

Given c_X, X was generated from the NHM with $(\mu_1, \eta_1) = (2, 1/0.25)$, $(\mu_2, \eta_2) = (4, 1/0.25)$ and γ_i $(i = 1, 2, \ldots, p)$ were independently generated by $1/(\text{Unif}(0.2, 1.5))^2$. Given c_A, A was generated under the setup of within-community connection probability 0.1 and between-community connection probability 0.06. For this scenario, we also introduced contamination into both X and A by the same way in Section 3.1.

In Figure 2, the ARIs for the integrative model, the NHM and the SBM were calculated by comparing their estimated labels with the overall cluster structure c. When estimating under integrative model, we set $K = 4$; under the NHM and SBM, we set $K = 2$. Similar to what we observe in Figure 1, the integrative model is better at capturing the overall clustering

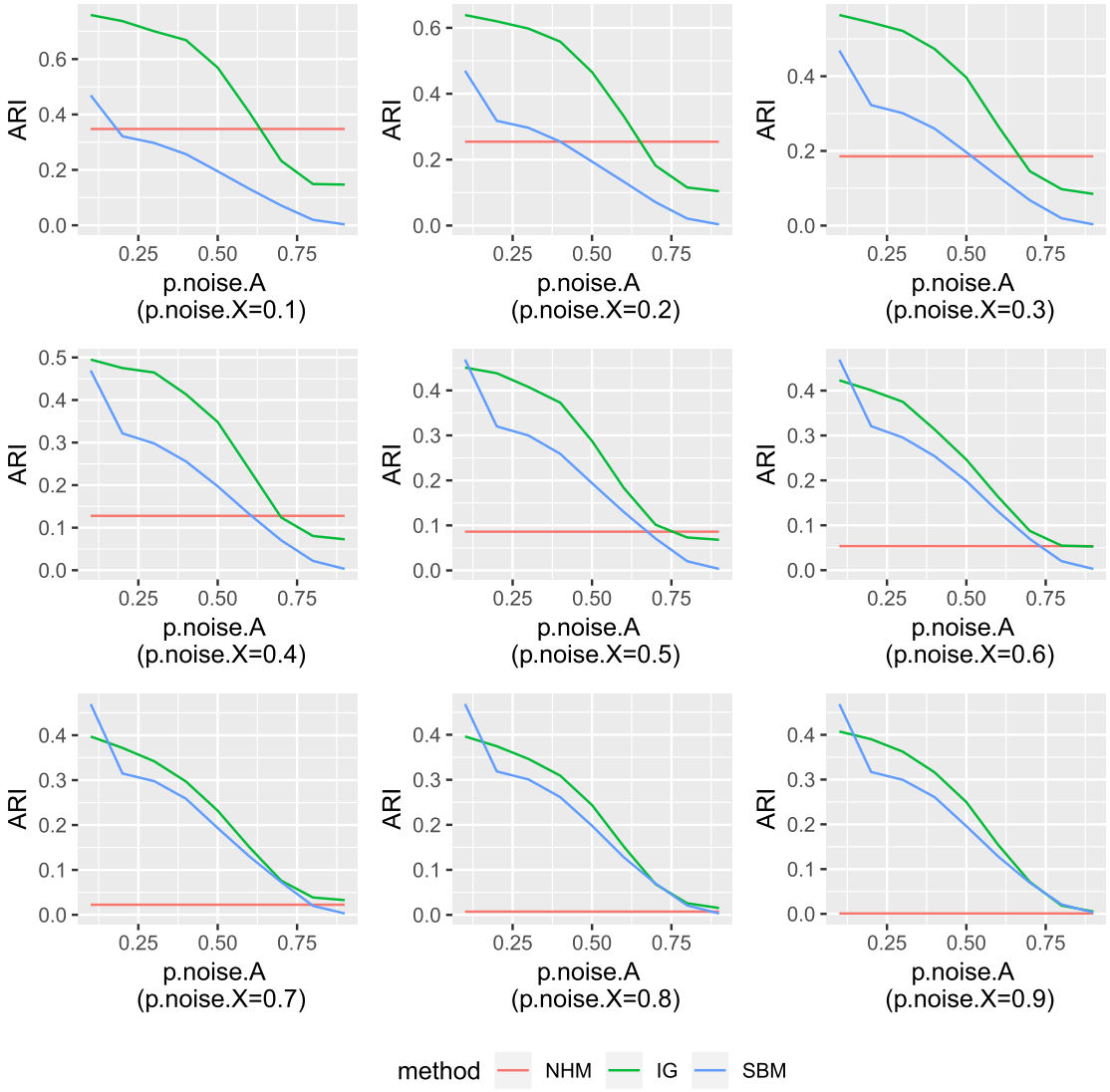

Figure 2: Comparison of the performance of integrative method vs SBM along vs NHM alone when X and A have independent community structures: the ARI between true/estimated and estimated labels as a function of $p.noise.X$ and $p.noise.A$.

structure than the separate models unless the contamination level is very high in one source and low in the other source, in which case the source with less contamination alone is the best choice. When both contamination levels in X and A are high, the performance of the integrative method tends to overlap with that of the SBM method on data source A, possibly due to the fact that $IG(x_i, b_i; \theta_k, \lambda_k)$ is dominated by $f_A(b_i|\lambda_k, c_i = k)$ in the calculation of posterior probabilities.

3.3 BIC-Type Criteria to Choose K

The BIC-type model selection criteria have been proposed to select the number of communities in networks, e.g., BIC (Saldana et al., 2017) and CBIC (Hu et al., 2020). Additional simulations

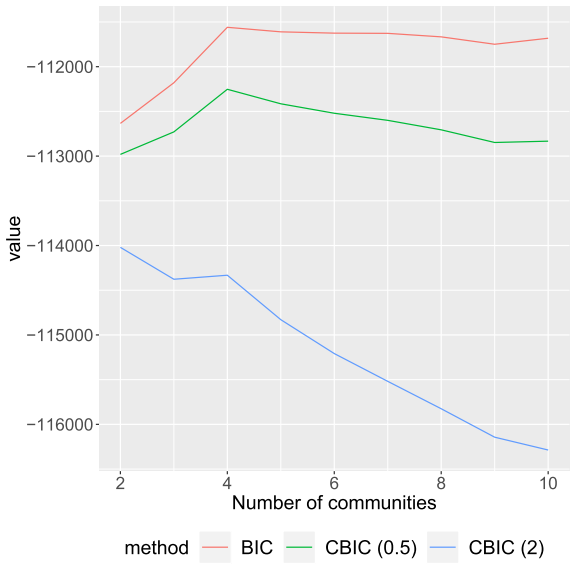

Figure 3: Values of BIC and CBIC over K in integrative network analysis.

were run to examine whether the BIC-type criteria are appropriate for integrative network analysis. We generated X and A with the true community number $K = 4$ and $p.noise.A = 0.2$, $p.noise.X = 0.2$. The rest of the parameter setting was identical to Section 3.1. We fitted the integrative model with $K = 2, \ldots, 10$, and used the three BIC-type criteria – BIC, CBIC with $\lambda = 0.5$, and CBIC with $\lambda = 2$ for model selection. Here we followed the definition of BIC and CBIC in Hu et al. (2020). That is, the three criteria are defined in the form of $2(\log \text{likelihood}) - \text{penalty}$. The simulation was repeated 100 times. The three lines in Figure 3 show the average values of BIC, CBIC with $\lambda = 0.5$, and CBIC with $\lambda = 2$. Both BIC and CBIC with $\lambda = 0.5$ reached their maximum at $\hat{K} = 4$ in all 100 replicates while CBIC with $\lambda = 2$ picked $\hat{K} = 2$ in all 100 replicates. It can be seen that CBIC with $\lambda = 0.5$ gives the most clear upside-down U-shape.

3.4 Robustness to Correlated Data Sources

Two correlated data sets $X^{(1)}$ and $X^{(2)}$ were generated under the NHM following the setup with $K = 3$, where the correlation coefficients between $X^{(1)}_{ij}$ and $X^{(2)}_{ij}$ varied from 0 to 0.8 with increments of 0.2. We added contamination to $X^{(1)}$ with $p.noise.X = 0.2$. Furthermore, we generated an adjacency matrix A with each entry A_{ij} being 1 with probability equal to the inverse distance in $X^{(2)}$, which were capped at 1. Finally, contamination was added to A with $p.noise.A = 0.4$. We then fitted the NHM on $X^{(1)}$, the SBM on A, and the integrative model on both. The simulation was repeated 200 times and the results were reported in Figure 4. There is no visible deteriorating performance as the correlation increases. Although correlated, A still provide extra information about the underlying community structure and the integrative analysis is always more accurate than either SBM on A alone or NHM alone, especially in the presence of contaminations in both data sources.

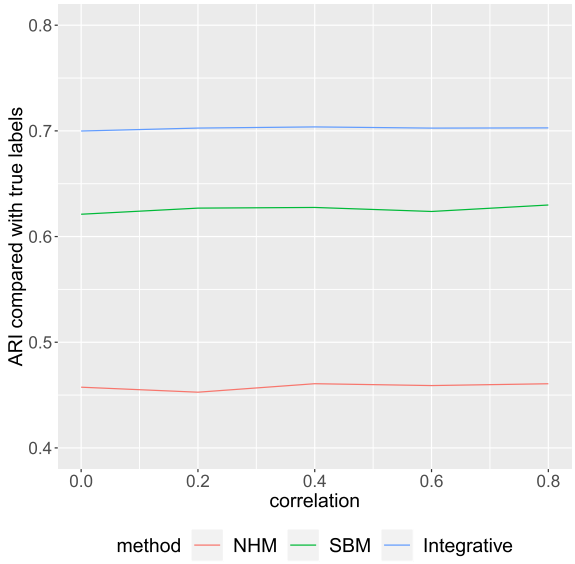

Figure 4: Comparison of the performance of integrative method vs SBM along vs NHM alone when X and A are correlated: the ARI between true and estimated labels as a function of correlation coefficients.

4 Clustering Analysis for Mouse Embryo Data

The integrative clustering method is applied to two independent data sets, both of which measure gene expressions in mouse embryonic cells. One is a microarray data set from Moliner et al. (2008) that measured gene expressions in bulk mouse embryonic stem (ES) cells. The other one is the scRNA sequence data in individual mouse ES cells from Islam et al. (2011). Cell samples in these two independent studies were cultivated following the same protocol (Andäng et al., 2008) and they were used as benchmark data sets sharing the same biomarker information in many bioinformatics papers (Wang et al., 2019; Di et al., 2011). The original bulk and single cell data were merged by gene names and genes that dropped out in more than 40 out of a total of 92 cells in the single cell data are deleted. After the pre-processing, there are 1015 genes shared by the microarray data and the scRNA data. Row data counts from both datasets were transformed using log-count-per-million (logCPM).

Both microarray and scRNA sequencing are high-throughput techniques that measure thousands to millions of genes simultaneously. Single cell expression data measure gene-specific transcript counts in individual cells, one cell at a time. Microarray data are gene expression levels from different samples, where each sample is a collection of cells. If all cells in a microarray sample have the same expression pattern, the two data sources would have the same underlying community structure/gene pathways. In this case, including both datasets in an integrative clustering analysis would lead to increased sample size and higher accuracy, especially given the extra noisy scRNA data. However, usually cells are heterogeneous with different expression patterns and scRNA data would provide information on individual cell level correlations between genes besides their mean levels. At the same time, scRNA has low data quality due to the limitation in sequencing a tiny amount of transcripts in a single cell. On the contrary, microarray data measures the average expression with high quality. Therefore, microarray data provide information on the clustering structure of the mean expression levels. The clustering

Table 2: Comparison of binary adjacency matrices using different thresholds.

	Threshold			
	0.024	0.022	0.021	0.020
Graph density	0.050	0.100	0.150	0.200
Signed R^2	0.669	0.495	0.236	0.058
ARI with X	0.290	0.283	0.239	0.247

structure on means (similar mean levels) and that on correlations (simultaneous expression) may not always agree. They can be viewed as complementary information on gene pathways where genes interact and collaborate to complete specific biological processes.

Microarray data are often modeled as normally distributed but the scRNA data are far from normal due to the extra noise in sequencing single cells (dropout, over-dispersion, batch effects, outliers, heterogeneity across cells, etc.). We ran additional simulations to examine the performance of the SBM on binary adjacent matrices converted from data generated under the NHM, which is robust and achieves similar ARI as NHM on the original data. Therefore, we treated the bulk cell microarray data as the multivariate normal X and summarized scRNA data into a binary adjacency matrix A. We constructed an affinity matrix based on the scRNA data where the affinity value was calculated as the inverse of the Euclidean distance between two genes. The range in the affinity matrix was further normalized into [0,1] through rescaling. The affinity matrix was further converted into a binary adjacency matrix where links with affinity values higher than 0.024 were set to be one and zero otherwise. This threshold of affinity value was chosen based on overall considerations of scale free topology fitting index signed R^2 (Zhang and Horvath, 2005), average link density and ARI with the microarray data X. Our choice of threshold 0.024 corresponds to the adjacency matrix with density approximately 0.05. Additionally, we tried three other thresholds 0.022, 0.021, and 0.020, corresponding to graph densities approximately 0.10, 0.15, and 0.20. The threshold 0.024 gave an adjacent matrix with the highest scale-free topology fitting index signed R^2 and the highest ARI with the microarray data X (Table 2).

We fitted the SBM on the scRNA adjacency matrix (A) and the NHM on the normalized bulk cell microarray data (X). The clusters from spectral clustering on A served as the initial labels in the pseudo-likelihood method for the SBM. The parameter estimates based on clusters from k-means on X were used as initial values for the EM algorithm of the NHM. The two sets of estimated labels from SBM on A only and NHM on X only were compared using ARI in Figure 5A. The figure shows that the ARI between NHM from X and the SBM from A decreases when the K increases. The ARI between the estimated clusters from the two data sources is larger than the reference threshold values in Table 1 under the corresponding K, suggesting potential benefits to conduct the integrative clustering method pooling information from both X and A.

We found a set of initial labels for the integrative model by combining the clustering structures from both data sources. An affinity matrix, denoted as aff(X), was produced from X by using the same inverse Euclidean transformation as we did in the scRNA data. Spectral clustering was conducted on aff(X) + A and the resulted labels were used as the initial labels in the integrative model. BIC and CBIC with $\lambda = 0.5$ and 2 were calculated as criteria to choose the number of clusters. The BIC and CBIC values of the group labels estimated from the integrative model versus the number of clusters K are plotted in Figure 5B. From the figure, the BIC and

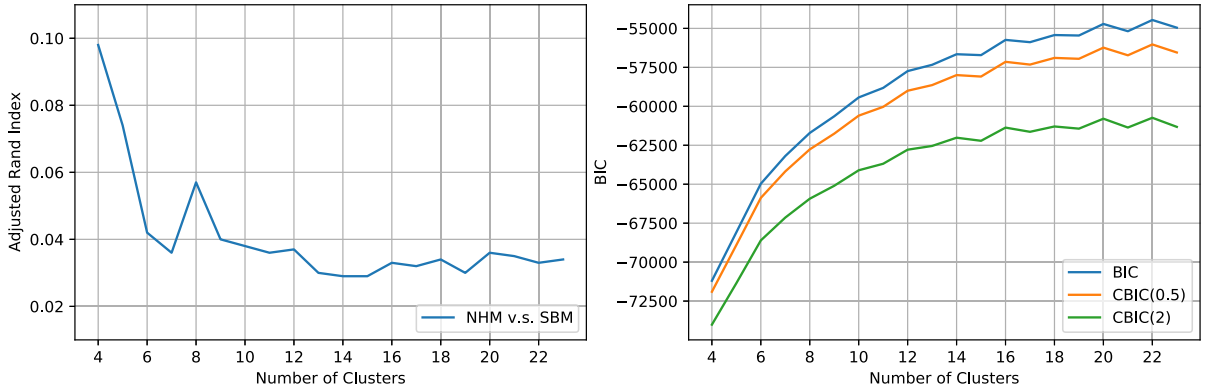

Figure 5: A (left panel): The ARI between estimated labels from X and A as a function of the number of clusters K; B (right panel): The BIC and CBIC ($\lambda = 0.5, 2$) as a function of the number of clusters K.

CBICs generally increase as the number of clusters K gets larger. The BIC (blue line) is always the highest followed by CBIC with $\lambda = 0.5$, due to the different penalty terms associated with K. We chose $K = 14$ because after this point all lines had their first drop. The clusters identified by the proposed integrative method as well as the results from single data sources were compared with known gene sets in MSigDB (https://www.gsea-msigdb.org/gsea/msigdb) to screen for gene sets with significant overlap. The three sets of results (known gene sets with significant overlap) using p-values threshold of 10^{-20} are exhibited in Table 3, Table 4 and Table 5 for the integrative model, the NHM and the SBM, respectively.

There are 32 gene sets in Table 3, 31 in Table 4, and 35 in Table 5. The last column is the cluster id that overlaps with the corresponding gene set. From the integrative clustering model, cluster 10 overlaps with 26 gene sets (p-values $< 10^{-20}$). Similarly, cluster 5 identified by the NHM on the bulk cell microarray data and cluster 3 by the SBM on the scRNA data also overlap with 26 and 28 gene sets, respectively. These three clusters estimated from the three models include 65, 60 and 65 genes respectively and share 31 genes in common. Moreover, 24 out of 26 common gene sets from Table 4 and Table 5, where most of them involve crucial functions in the early developments of an organism, were also identified by the integrative method in Table 3. Four gene sets discovered by the SBM in Table 5 but missed by the NHM in Table 4 are actually reported by the integrative model in Table 3. Two gene sets that are uniquely discovered by the integrative method, did not reach the p-value threshold of 10^{-20} in the methods using one data set alone (p-values $= 3.65 \times 10^{-7}$, 8.43×10^{-10} in X and 4.97×10^{-11}, 2.03×10^{-19} in A).

5 Discussion and Future Research Directions

We proposed an integrative clustering method that combines the NHM for microarray data and the SBM for a binary adjacency matrix derived from scRNA data. An EM algorithm was developed following the spirit of the pseudo-likelihood method (Amini et al., 2013). In real applications, since researchers often do not know whether the underlying community structures in different data sources agree, a challenging question is whether integrative method provides higher accuracy than separate clustering analysis on individual data sources. We proposed an

Table 3: GSEA for gene clusters identified by the integrative clustering model.

Gene set name	p-value	FDR q-value	K
REACTOME NONSENSE MEDIATED DECAY NMD INDEPENDENT OF THE EXON JUNCTION COMPLEX EJC	1.90×10^{-68}	1.04×10^{-64}	10
HSIAO HOUSEKEEPING GENES	1.13×10^{-67}	3.11×10^{-64}	10
KEGG RIBOSOME	8.85×10^{-67}	1.62×10^{-63}	10
REACTOME SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING TO MEMBRANE	5.85×10^{-66}	8.05×10^{-63}	10
REACTOME NONSENSE MEDIATED DECAY NMD	1.45×10^{-65}	1.59×10^{-62}	10
REACTOME SELENOAMINO ACID METABOLISM	2.61×10^{-65}	2.39×10^{-62}	10
REACTOME EUKARYOTIC TRANSLATION INITIATION	4.65×10^{-65}	3.65×10^{-62}	10
REACTOME INFLUENZA INFECTION	5.83×10^{-64}	4.01×10^{-61}	10
REACTOME REGULATION OF EXPRESSION OF SLITS AND ROBOS	7.99×10^{-60}	4.88×10^{-57}	10
REACTOME RRNA PROCESSING IN THE NUCLEUS AND CYTOSOL	3.42×10^{-58}	1.88×10^{-55}	10
REACTOME RRNA PROCESSING	2.40×10^{-57}	1.20×10^{-54}	10
REACTOME SIGNALING BY ROBO RECEPTORS	1.74×10^{-56}	7.97×10^{-54}	10
REACTOME INFECTIOUS DISEASE	8.35×10^{-56}	3.53×10^{-53}	10
REACTOME TRANSLATION	9.95×10^{-55}	3.91×10^{-52}	10
REACTOME AXON GUIDANCE	1.50×10^{-52}	5.49×10^{-50}	10
REACTOME METABOLISM OF AMINO ACIDS AND DERIVATIVES	3.63×10^{-49}	1.25×10^{-46}	10
REACTOME DISEASE	1.05×10^{-48}	3.38×10^{-46}	10
REACTOME DEVELOPMENTAL BIOLOGY	3.18×10^{-42}	9.71×10^{-40}	10
REACTOME METABOLISM OF RNA	2.78×10^{-41}	7.85×10^{-39}	10
BILANGES SERUM AND RAPAMYCIN SENSITIVE GENES	2.86×10^{-41}	7.85×10^{-39}	10
DIAZ CHRONIC MEYLOGENOUS LEUKEMIA UP	3.06×10^{-36}	1.68×10^{-32}	13
OSMAN BLADDER CANCER DN	6.46×10^{-31}	1.69×10^{-28}	10
LOPEZ MBD TARGETS	3.04×10^{-27}	8.36×10^{-24}	2
REACTOME ACTIVATION OF THE MRNA UPON BINDING OF THE CAP BINDING COMPLEX AND EIFS AND SUBSEQUENT BINDING TO 43S	3.73×10^{-29}	9.34×10^{-27}	10
WANG TUMOR INVASIVENESS UP	1.59×10^{-28}	3.79×10^{-26}	10
PECE MAMMARY STEM CELL UP	4.97×10^{-28}	1.14×10^{-25}	10
CHNG MULTIPLE MYELOMA HYPERPLOID UP	2.68×10^{-27}	5.89×10^{-25}	10
REN ALVEOLAR RHABDOMYOSARCOMA DN	9.65×10^{-26}	5.31×10^{-22}	4
PUJANA BRCA1 PCC NETWORK	4.53×10^{-25}	1.24×10^{-21}	13
TIEN INTESTINE PROBIOTICS 6HR UP	7.00×10^{-25}	1.48×10^{-22}	10
PILON KLF1 TARGETS DN	7.62×10^{-24}	1.40×10^{-20}	13
GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	2.20×10^{-21}	3.03×10^{-18}	13

Table 4: GSEA for gene clusters identified by the NHM.

Gene set name	p-value	FDR q-value	K
KEGG RIBOSOME	6.53×10^{-66}	3.59×10^{-62}	5
REACTOME NONSENSE MEDIATED DECAY NMD INDEPENDENT OF THE EXON JUNCTION COMPLEX EJC	1.17×10^{-64}	3.21×10^{-61}	5
REACTOME SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING TO MEMBRANE	2.30×10^{-62}	4.22×10^{-59}	5
REACTOME REGULATION OF EXPRESSION OF SLITS AND ROBOS	3.28×10^{-62}	4.51×10^{-59}	5
REACTOME NONSENSE MEDIATED DECAY NMD	5.30×10^{-62}	5.84×10^{-59}	5
REACTOME SELENOAMINO ACID METABOLISM	9.14×10^{-62}	8.38×10^{-59}	5
REACTOME EUKARYOTIC TRANSLATION INITIATION	1.56×10^{-61}	1.22×10^{-58}	5
REACTOME SIGNALING BY ROBO RECEPTORS	7.18×10^{-59}	4.94×10^{-56}	5
HSIAO HOUSEKEEPING GENES	4.66×10^{-58}	2.85×10^{-55}	5
REACTOME INFLUENZA INFECTION	5.54×10^{-58}	3.05×10^{-55}	5
REACTOME RRNA PROCESSING IN THE NUCLEUS AND CYTOSOL	3.60×10^{-55}	1.78×10^{-52}	5
REACTOME RRNA PROCESSING	2.18×10^{-54}	9.99×10^{-52}	5
REACTOME INFECTIOUS DISEASE	1.26×10^{-53}	5.35×10^{-51}	5
REACTOME AXON GUIDANCE	3.66×10^{-53}	1.44×10^{-50}	5
REACTOME TRANSLATION	1.03×10^{-49}	3.80×10^{-47}	5
REACTOME METABOLISM OF AMINO ACIDS AND DERIVATIVES	8.81×10^{-47}	3.03×10^{-44}	5
REACTOME METABOLISM OF RNA	1.22×10^{-43}	3.95×10^{-41}	5
REACTOME DEVELOPMENTAL BIOLOGY	4.00×10^{-43}	1.22×10^{-40}	5
REACTOME DISEASE	1.93×10^{-41}	5.59×10^{-39}	5
BILANGES SERUM AND RAPAMYCIN SENSITIVE GENES	7.92×10^{-34}	2.18×10^{-31}	5
REN ALVEOLAR RHABDOMYOSARCOMA DN	7.85×10^{-32}	4.32×10^{-28}	1
REACTOME ACTIVATION OF THE MRNA UPON BINDING OF THE CAP BINDING COMPLEX AND EIFS AND SUBSEQUENT BINDING TO 43S	5.8×10^{-30}	1.52×10^{-27}	5
MARTENS TRETINOIN RESPONSE DN	3.06×10^{-29}	7.65×10^{-27}	5
KIM BIPOLAR DISORDER OLIGODENDROCYTE DENSITY CORR UP	1.38×10^{-28}	3.79×10^{-25}	14
TIEN INTESTINE PROBIOTICS 6HR UP	1.49×10^{-25}	3.56×10^{-23}	5
KIM ALL DISORDERS OLIGODENDROCYTE NUMBER CORR UP	1.76×10^{-24}	3.22×10^{-21}	14
WANG TUMOR INVASIVENESS UP	5.74×10^{-24}	1.32×10^{-21}	5
PASINI SUZ12 TARGETS DN	2.76×10^{-23}	1.52×10^{-19}	8
DIAZ CHRONIC MEYLOGENOUS LEUKEMIA UP	1.02×10^{-22}	5.62×10^{-19}	3
BILANGES SERUM RESPONSE TRANSLATION	4.47×10^{-21}	9.84×10^{-19}	5
OSMAN BLADDER CANCER DN	5.04×10^{-21}	1.07×10^{-18}	5

Table 5: GSEA for gene clusters identified by the SBM.

Gene set name	p-value	FDR q-value	K
REACTOME NONSENSE MEDIATED DECAY NMD INDEPENDENT OF THE EXON JUNCTION COMPLEX EJC	1.90×10^{-68}	1.04×10^{-64}	3
HSIAO HOUSEKEEPING GENES	1.13×10^{-67}	3.11×10^{-64}	3
KEGG RIBOSOME	8.85×10^{-67}	1.62×10^{-63}	3
REACTOME SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING TO MEMBRANE	5.85×10^{-66}	8.05×10^{-63}	3
REACTOME NONSENSE MEDIATED DECAY NMD	1.45×10^{-65}	1.59×10^{-62}	3
REACTOME SELENOAMINO ACID METABOLISM	2.61×10^{-65}	2.39×10^{-62}	3
REACTOME EUKARYOTIC TRANSLATION INITIATION	4.65×10^{-65}	3.65×10^{-62}	3
REACTOME INFLUENZA INFECTION	5.83×10^{-64}	4.01×10^{-61}	3
REACTOME REGULATION OF EXPRESSION OF SLITS AND ROBOS	7.99×10^{-60}	4.88×10^{-57}	3
REACTOME RRNA PROCESSING IN THE NUCLEUS AND CYTOSOL	3.42×10^{-58}	1.88×10^{-55}	3
REACTOME RRNA PROCESSING	2.40×10^{-57}	1.20×10^{-54}	3
REACTOME SIGNALING BY ROBO RECEPTORS	1.74×10^{-56}	7.97×10^{-54}	3
REACTOME INFECTIOUS DISEASE	8.35×10^{-56}	3.53×10^{-53}	3
REACTOME TRANSLATION	9.95×10^{-55}	3.91×10^{-52}	3
REACTOME AXON GUIDANCE	1.50×10^{-52}	5.49×10^{-50}	3
REACTOME METABOLISM OF AMINO ACIDS AND DERIVATIVES	3.63×10^{-49}	1.25×10^{-46}	3
REACTOME DISEASE	1.05×10^{-48}	3.38×10^{-46}	3
REACTOME DEVELOPMENTAL BIOLOGY	3.18×10^{-42}	9.71×10^{-40}	3
REACTOME METABOLISM OF RNA	2.78×10^{-41}	7.85×10^{-39}	3
BILANGES SERUM AND RAPAMYCIN SENSITIVE GENES	2.86×10^{-41}	7.85×10^{-39}	3
DIAZ CHRONIC MEYLOGENOUS LEUKEMIA UP	5.33×10^{-35}	2.93×10^{-31}	2
OSMAN BLADDER CANCER DN	6.46×10^{-31}	1.69×10^{-28}	3
REN ALVEOLAR RHABDOMYOSARCOMA DN	1.60×10^{-29}	8.78×10^{-26}	9
REACTOME ACTIVATION OF THE MRNA UPON BINDING OF THE CAP BINDING COMPLEX AND EIFS AND SUBSEQUENT BINDING TO 43S	3.73×10^{-29}	9.34×10^{-27}	3
PUJANA BRCA1 PCC NETWORK	1.07×10^{-28}	2.96×10^{-25}	2
WANG TUMOR INVASIVENESS UP	1.59×10^{-28}	3.79×10^{-26}	3
PECE MAMMARY STEM CELL UP	4.97×10^{-28}	1.14×10^{-25}	3
CHNG MULTIPLE MYELOMA HYPERPLOID UP	2.68×10^{-27}	5.89×10^{-25}	3
LOPEZ MBD TARGETS	1.77×10^{-23}	4.86×10^{-20}	8
TIEN INTESTINE PROBIOTICS 6HR UP	7.00×10^{-25}	1.48×10^{-22}	3
TIEN INTESTINE PROBIOTICS 24HR DN	7.45×10^{-23}	1.52×10^{-20}	3
BLALOCK ALZHEIMERS DISEASE DN	1.04×10^{-22}	1.43×10^{-19}	2
JISON SICKLE CELL DISEASE DN	8.01×10^{-22}	1.57×10^{-19}	3
KIM BIPOLAR DISORDER OLIGODENDROCYTE DENSITY CORR UP	1.18×10^{-21}	1.63×10^{-18}	8
PASINI SUZ12 TARGETS DN	1.35×10^{-21}	2.47×10^{-18}	9

empirical guideline to this question. Furthermore, we also investigated the performance of the integrative model on data sources with different community structures. In that case, we considered the intersection of the community structures in different data sources to be the overall community structure, and different data sources provide complementary information on the overall community structure. Therefore, the integrative methods are also helpful in disclosing overlapping community structures.

We plan to conduct future research in the following directions. We provided an empirical guideline for the choice between the integrative method and separate analysis. The thresholds, which are determined by simulations studies, are ARIs between the estimated clusters from different data sources when the integrative method starts to outperform methods using a single data source. Future research may derive the asymptotic distributions of ARI under the null hypothesis that the community structures from different data sources are identical. Furthermore, if there is biological or clinical belief that the two related data sources describe the same disease or biophysical processes, it is reasonable to carry out integrative clustering analysis without checking the ARI. Moreover, the integrative method can also be used as an exploratory data analysis tool to discover new clusters. The proposed integrative clustering method falls into the probability clustering category where the log-likelihood for each dataset are summed and maximized jointly. Therefore, models other than NHM or SBM can be assumed as long as a likelihood, quasi-likelihood or pseudo-likelihood can be written out. Furthermore, in cases with more than two datasets available, their log-likelihoods can still be summed but the verification of common clustering structure can no longer be judged by pairwise ARI. Testing procedures or empirical guidelines need to be developed to examine whether clustering structures from more than two sources are the same or not. Finally, although both NHM and SBM are based on expression data, they extract information from different aspects (i.e. means and Euclidean distances), which may lead to different sample sizes and different contributions in the joint likelihood. The sizes of the two likelihood functions may not be comparable in the E-step when calculating the posterior probabilities. It is possible that the group structure are dominated by either the NHM or SBM model because of this imbalance. Therefore, when one data source is more important or reliable than the other, one may weigh the data sets accordingly in the joint likelihood.

Supplementary Material

Code for the integrative analysis and the data used in the real data analysis are available at https://github.com/yangliuqing1992/Integrative_clustering.

References

Abbe E (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(177): 1–86.

Amini AA, Chen A, Bickel PJ, Levina E (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4): 2097–2122.

Andäng M, Moliner A, Doege CA, Ibañez CF, Ernfors P (2008). Optimized mouse ES cell culture system by suspension growth in a fully defined medium. *Nature Protocols*, 3(6): 1013–1017.

Bickel PJ, Chen A (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50): 21068–21073.

Di Y, Schafer DW, Cumbie JS, Chang JH (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1): 1–28.

Fletcher R (2013). *Practical Methods of Optimization*. John Wiley & Sons.

Forcato M, Romano O, Bicciato S (2021). Computational methods for the integrative analysis of single-cell data. *Briefings in Bioinformatics*, 22(3): 1–10.

Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2): 129–233.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4): 929–944.

Holland PW, Laskey KB, Leinhardt S (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2): 109–137.

Hu J, Qin H, Yan T, Zhao Y (2020). Corrected bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, 115(532): 1771–1783.

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7): 1160–1167.

Karrer B, Newman ME (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1): 016107.

Lock EF, Dunson DB (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20): 2610–2616.

Lock EF, Hoadley KA, Marron JS, Nobel AB (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1): 523–542.

Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11): 4245–4250.

Moliner A, Enfors P, Ibáñez CF, Andäng M (2008). Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. *Stem Cells and Development*, 17(2): 233–243.

Morris CN, Lysy M (2012). Shrinkage estimation in multilevel normal models. *Statistical Science*, 27(1): 115–134.

Newman ME, Clauset A (2016). Structure and inference in annotated networks. *Nature Communications*, 7(1): 1–11.

Nguyen T, Tagett R, Diaz D, Draghici S (2017). A novel approach for data integration and disease subtyping. *Genome Research*, 27(12): 2025–2039.

Perkins AD, Langston MA (2009). Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinformatics*, 10: 1–11.

Priness I, Maimon O, Ben-Gal I (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8(1): 1–12.

Rand WM (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Rappoport N, Shamir R (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic acids research*, 46(20): 10546–10562.

Saldana DF, Yu Y, Feng Y (2017). How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1): 171–181.

Thompson A, May MR, Moore BR, Kopp A (2020). A hierarchical bayesian mixture model for

inferring the expression state of genes in transcriptomes. *Proceedings of the National Academy of Sciences*, 117(32): 19339–19346.

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3): 333–337.

Wang H, Nie F, Huang H (2013). Multi-view clustering and feature learning via structured sparsity. In: Dasgupta S, McAllester D (eds.) *International Conference on Machine Learning*, 352–360.

Wang T, Li B, Nelson CE, Nabavi S (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, 20(1): 1–16.

Wu D, Wang D, Zhang MQ, Gu J (2015). Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification. *BMC Genomics*, 16(1): 1022.

Xu Z, Ke Y, Wang Y, Cheng H, Cheng J (2012). A model-based approach to attributed graph clustering. In: Dasgupta S, McAllester D (eds.) *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 505–516.

Yan B, Sarkar P (2021). Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 116(534): 734–745.

Zhang B, Horvath S (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1): 17.

Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19): 9379–9391.

Zhao Y (2017). A survey on theoretical advances of community detection in networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5): e1403.

Zhao Y, Levina E, Zhu J (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4): 2266–2292.