Computational and Structural Biotechnology Journal 18 (2020) 2583–2595

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Joint learning of multiple gene networks from single-cell gene expression data

Nuosi Wu [a], Fu Yin [a], Le Ou-Yang [a,b,c,*], Zexuan Zhu [d,*], Weixin Xie [a]

[a] College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
[b] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, and Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ), Shenzhen University, Shenzhen, China
[c] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[d] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

ARTICLE INFO

Article history:
Received 7 May 2020
Received in revised form 31 August 2020
Accepted 1 September 2020
Available online 10 September 2020

Keywords:
Single-cell RNA sequencing
Gene network
Graphical model

ABSTRACT

Inferring gene networks from gene expression data is important for understanding functional organizations within cells. With the accumulation of single-cell RNA sequencing (scRNA-seq) data, it is possible to infer gene networks at single cell level. However, due to the characteristics of scRNA-seq data, such as cellular heterogeneity and high sparsity caused by dropout events, traditional network inference methods may not be suitable for scRNA-seq data. In this study, we introduce a novel joint Gaussian copula graphical model (JGCGM) to jointly estimate multiple gene networks for multiple cell subgroups from scRNA-seq data. Our model can deal with non-Gaussian data with missing values, and identify the common and unique network structures of multiple cell subgroups, which is suitable for scRNA-seq data. Extensive experiments on synthetic data demonstrate that our proposed model outperforms other compared state-of-the-art network inference models. We apply our model to real scRNA-seq data sets to infer gene networks of different cell subgroups. Hub genes in the estimated gene networks are found to be biological significance.

1. Introduction

Gene expression and regulation are the foundation of biological processes. Inferring the regulatory relationships between genes could help to understand the functional organizations within cells systematically. With the rapid development of high-throughput techniques such as microarray and RNA sequencing, it becomes possible to infer gene networks in genome-scale. A wide range of computational approaches have been developed to infer gene networks from gene expression data, including methods based on Boolean models [1], ordinary differential equations [2], Bayesian approaches [3], Gaussian graphical models [4], regression [5] and mutual information [6]. These methods have been successfully used to deal with gene expression data collected by bulk sequencing technologies, which perform high-throughput sequencing on a population of millions of cells and output the average expression levels of genes. Therefore, the heterogeneous information of different cell types is obscured in bulk gene expression data.

Recently, the emergence and development of single-cell experimental techniques allow us to quantify gene expression at single-cell resolution. Microfluidics techniques and combinatorial indexing strategies made it possible to sequencing thousands of cells in one experiment. Nevertheless, the inherent characteristics of single-cell RNA sequencing (scRNA-seq) data have not been changed. For example, due to the small amount of DNA in a single cell (6 pg in total for human [7]), it is necessary to amplify nucleic acids a few hundred times for sequencers to capture the signals. But the amplification is prone to bring in biased information, leading to unreliable results in downstream analysis. Although methods designed for bulk sample data could be applied to single-cell data directly, these methods may result in biased estimations due to the challenges derived from single-cell RNA sequencing (scRNA-seq) [8].

A key challenge in inferring gene networks from scRNA-seq data is to deal with the large fraction of observed “0”s in the data [9]. Some of these “0”s are true “0”s, whereas the rest are false “0”s caused by technical limitations (also called “dropout events”) [10]. To deal with dropout events, some imputation methods have been developed to recover original signals by imputing the false

[*] Corresponding authors.
E-mail addresses: leouyang@szu.edu.cn (L. Ou-Yang), zhuzx@szu.edu.cn (Z. Zhu).

https://doi.org/10.1016/j.csbj.2020.09.004

“0”s in the gene expression matrices [11–13]. However, as pointed out by Andrews et al. [14] and Chen et al. [15], due to the high rate of dropout events, imputation methods may distort the overall shape of the gene expression distribution and reduce the repeatability of cell type specific markers, which may induce bias in downstream analysis such as network estimation [9,16]. Network inference across cell types with imputed data also suffers from homogenization which may reduce the differences between different cell types. Furthermore, imputation methods relying only on the observed gene expression data may strengthen the intrinsic signal contained in the observed data, and introduce circularity that can generate inflated false-positive results in network inference [17].

The cellular heterogeneity also generates new challenges for inferring gene networks from scRNA-seq data. A special case is to infer the differential network between samples collected from two distinct groups. Similar to the methods developed for microarray data, Wang et al. [18] measured the connectivity between groups with mean absolute distance, and used a permutation approach to check the significance of differential edges. Chiu et al. [19] compared the correlation coefficients among groups and the statistical significance was tested by a cumulative distribution function. Fisher transformation was applied in their study so that the sample size related biases was eliminated. Dai et al. [20] studied gene-gene interactions for every single cell, which could provide new perspectives for clustering and pseudo-trajectory. In general, scRNA-seq data collected in one experiment may contain cells belong to more than two cell subgroups. As the cells belong to different subgroups are originated from the same tissue, their networks may share some common structures, and inferring the gene network of each cell subgroup separately may not be able to make use of the similarities between different cell subgroups. Thus, joint estimation of multiple gene networks, which can draw support from multiple cell subgroups, may lead to more accurate estimation of gene networks [21,22].

Gaussian graphical models (GGM) have been widely used in inferring gene networks from microarray data. Based on Gaussian graphical models, the estimation of a gene network can be achieved by estimating the inverse of the covariance matrix (also named as the precision matrix) of the corresponding multivariate Gaussian distribution. Based on the assumption that gene networks are sparse, Yuan and Lin [23] imposed a ℓ1-norm penalty on the non-diagonal elements of the precision matrix when estimating the precision matrix. Friedman et al. [24] proposed a graphical lasso model, which transforms the penalized log-likelihood optimization problem into a lasso regression problem. The CLIME estimation method proposed by Cai et al. [25] directly estimates each column of the precision matrix, which can easily parallelize

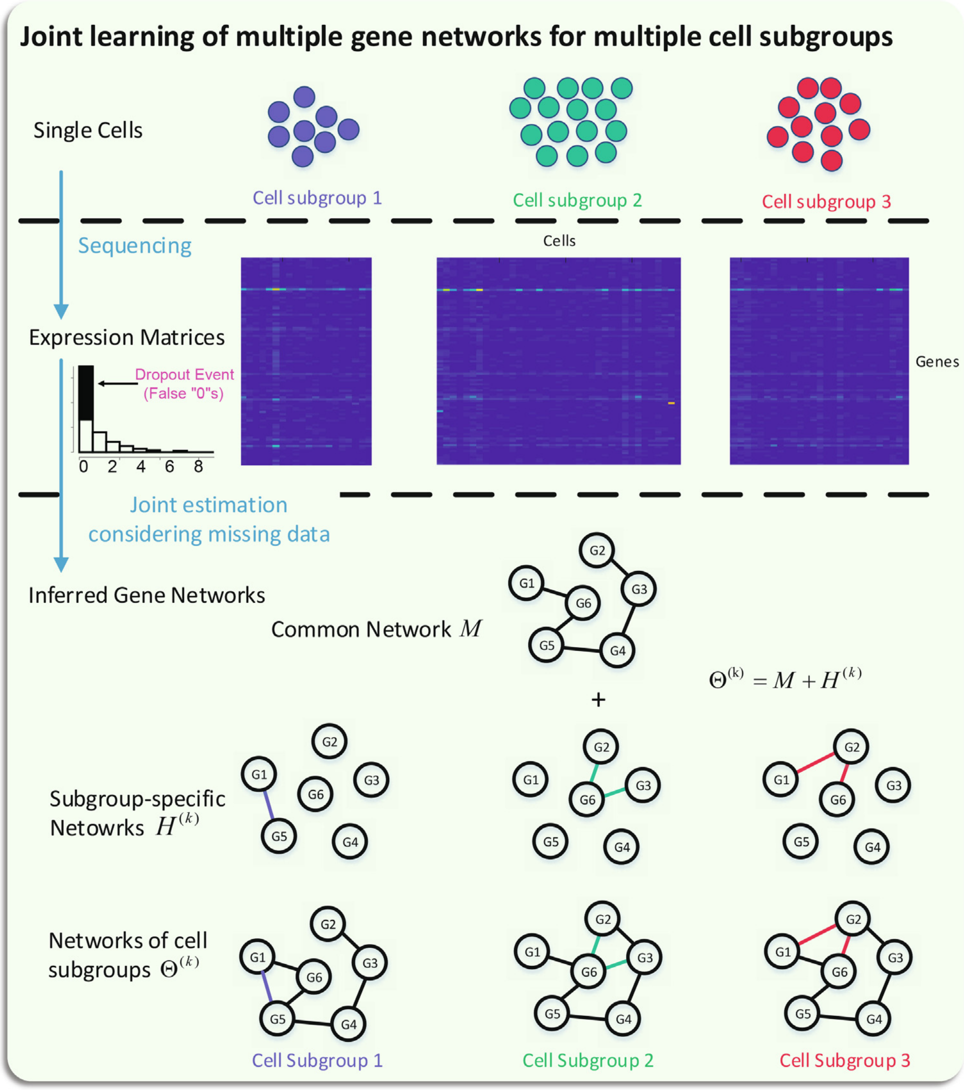

Fig. 1. The flowchart of the proposed Joint Gaussian Copula Graphical Model.

for large scale problem. For data sets with multiple states, joint graphical models have been proposed to estimate networks jointly [26,27]. For example, Guo et al. [28] used a hierarchical penalty to force different networks to have a similar sparse structure. Danaher et al. [29] imposed fused lasso and group lasso penalties to the Gaussian graphical model, and jointly estimate multiple graphical models. Zhang et al. [30] integrated gene expression data collected from multiple platforms and multiple conditions for joint network estimation.

However, Gaussian graphical models require that the data should follow a normal distribution, which limits their applications on scRNA-seq data [31], as RNA sequencing data can be considered to obey negative binomial or Poisson distribution [32,33]. In order to model the graph under non-Gaussian conditions, Liu et al. [34] proposed a nonparanormal distribution and introduced semiparametric Gaussian copula graphical models to infer the conditional dependence between random variables that do not follow normal distribution. Xue and Zou [35] proposed a rapid method for estimating the correlation coefficient matrix in nonparanormal distribution. Nevertheless, these joint graphical models are based on the assumption that the observed data are complete, which make it difficult to directly apply these models to scRNA-seq data with missing values.

In this paper, we propose a Joint Gaussian Copula Graphical Model (JGCGM) (Fig. 1) to jointly infer the gene networks of multiple cell subgroups from scRNA-seq data. Similar to existing joint graphical models [30,36], our model decomposes the gene network of each subgroup into two parts, i.e., a common part that shared across all cell subgroups and a subgroup-specific part that capturing the edges specific to each cell subgroup. To deal with the missing value included in scRNA-seq data, we extend our model to handle non-Gaussian data with missing values. The modified Kendall's tau proposed by Wang et al. [37] is employed for estimating the correlation coefficient matrix. We first evaluate the performance of our proposed model on synthetic data. Then we apply our model on real scRNA-seq data to infer gene networks of different cell subgroups. The hub genes in our estimated gene networks are closely related to cell differentiation.

In the remaining of this paper, we first introduce details of the proposed method JGCGM in Section 2. Then we compare JGCGM with other state-of-the-arts network inference approaches on synthetic data in Section 3. In Section 4, we apply our model on two real scRNA-seq datasets to demonstrate the effectiveness of our proposed model. Finally we conclude the paper in Section 5.

2. Methods

2.1. Gaussian graphical models

Assuming $Z = (Z_1, \ldots, Z_p)^T$ is a p-dimensional random vector that follows multivariate normal distribution $Z \sim N(\mu, \Sigma)$. Let $G = (V, E)$ denotes an undirected graph, where V represents the set of nodes (each node in V represents a random variable in Z) and E describes the edges between nodes in V. Then Z is said to satisfy the Gaussian graphical model with graph G, only when the following statements are equivalent [38,39]:

- $Z_i \perp\!\!\!\perp Z_j | Z_{\backslash\{i,j\}}$;
- $E_{ij} = 0$;
- $\Theta_{ij} = 0$.

Here $Z_{\backslash\{i,j\}}$ is defined as $(Z_s : 1 \leqslant s \leqslant p, s \neq i, j)$, and $\Theta = \Sigma^{-1}$ is referred to as precision matrix. The statements show that in Gaussian graphical model, variable Z_i is conditional independent with variable Z_j if and only if there is no connection between these two variables in graph G. Moreover, it is feasible to recover the graphical model by estimating the parameters in the corresponding precision matrix.

Reconstruction of the precision matrix is not a simple task, especially when the number of samples n is much smaller than the dimension of variables p. It is hard to obtain the precision matrix from sample covariance matrix directly since sample covariance matrix is usually not invertible. A common approach is to maximize the log-likelihood function and assume the precision matrix to be sparse [40]. The log-likelihood function is shown as follows:

$$\ell(\Theta) = -tr(S\Theta) + \log\det(\Theta) - \lambda\sum_{i\neq j}|\Theta_{ij}|. \quad (1)$$

where S is the sample covariance matrix, $\det(\cdot)$ denotes the determinant of a matrix, and λ is a non-negative tuning parameter to control the sparsity of the graph.

Eq. (1) just fit for homogeneous datasets. However, we are also faced with data from multiple groups. Suppose there are K groups of data, each of which follows a multivariate normal distribution $X^{(k)} \sim N\left(\mu^{(k)}, \Sigma^{(k)}\right), k = 1, \ldots K$. In order to get more accurate inference results, one could joint estimate the Gaussian graphical models by maximizing the following objective function:

$$\text{maximize}_{\{\Theta^{(k)}\}\succeq 0} \quad \sum_{k=1}^{K} n_k\left[\log\left\{\det\left(\Theta^{(k)}\right)\right\} - \text{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] - \underbrace{F\left(\left\{\Theta^{(k)}\right\}\right)}_{\text{penalty}}, \quad (2)$$

where $n_k$represents the number of samples from the k-th group. The penalty function in Eq. (2) is used to control the structure of the estimated networks, and could be modified into different specific forms in accordance with application scenarios.

2.2. Joint Gaussian copula graphical model for scRNA-seq data

We first introduce the nonparanormal distribution. Given a vector $X = (X_1, \ldots, X_p)^T$, we say that it follows nonparanormal distribution (written as $X \sim NPN(\mu, \Sigma, f)$) if there exists a set of monotonous and derivable functions $\{f_d\}, d = 1, \ldots, p$, such that:

$$Z = f(X) \sim N(\mu, \Sigma),$$

where $f(X) = (f_1(X_1), \ldots, f_p(X_p))^T$. The nonparanormal distribution is actually a Gaussian Copula. According to the previous studies [34,41,35,42], the sparsity pattern of $\Theta = \Sigma^{-1}$ encodes the conditional dependence between X.

Nonparanormal distribution provides a widely applicable model for skewed distributed data, and it can also be suitable for scRNA-seq. Due to the presence of cell subgroups (types), observations usually belong to different distributions. Suppose we have a set of scRNA-seq data X measuring the expression levels of p genes and composed of observations collected from K cell subgroups. Suppose each cell subgroup follows a nonparanormal distribution, i.e. $X^{(k)} \sim NPN\left(\mu^{(k)}, \Sigma^{(k)}, f^{(k)}\right), k = 1, \ldots, K$. Similar to the above Guassian copula graphical model for single distribution, joint estimation of gene networks for all cell subgroups can be transferred into solving problem (2).

Because the observations of K cell subgroups often come from the same tissue, it is reasonable to assume that there are some gene interactions shared across all cell subgroups. These gene interactions constitute a common network and reflects the similarities between different cell subgroups. Besides, the network corresponding to each subgroup may not be exactly the same, and the heterogeneity between cell subgroups will also be reflected in the network structure. These subgroup-specific edges make up a subgroup-specific network for each cell subgroup. In other words, the network $\Theta^{(k)}$ (represented by its precision matrix) for each cell subgroup can be regarded as the combination of the common network M and subgroup-specific network $H^{(k)}$, i.e., $\Theta^{(k)} = M + H^{(k)}$, where K cell subgroups share the same common network but have their unique subgroup-specific networks. Both the common network or the subgroup-specific networks are assumed to be sparse, while the sparsity may be different. To model the homogeneity and heterogeneity simultaneously for joint estimation of multiple networks, similar to [30], we propose a novel penalty function as follows:

$$F(\Theta) = \lambda\alpha K \sum_{i \neq j} |M_{ij}| + \lambda(1-\alpha) \sum_{k=1}^{K} \sum_{i \neq j} |H_{ij}^{(k)}|. \tag{3}$$

Here $\lambda > 0$ is the tuning parameter controlling the sparsity of all networks. $0 < \alpha < 1$ is the proportional factor, which controls the relative scales of M and $H^{(k)}$. It should be closer to 0 when the difference between the cell subgroups is not large, which means strong sparsity constraints are imposed on subgroup-specific networks, encouraging fewer edges. On the contrary, α should be larger if the heterogeneity among cell subgroups is more prominent.

By substitute (3) into (2), we obtain the final objective function of the joint Gaussian copula graph model (JGCGM) for scRNA-seq data:

$$\begin{aligned} \text{minimize}_{\{\Theta^{(k)}, H^{(k)}, M\}} \quad & \sum_{k=1}^{K} n_k \left[\text{tr}\left(S^{(k)} \Theta^{(k)} \right) - \log\left\{ \det\left(\Theta^{(k)} \right) \right\} \right] \\ & + \lambda\alpha K \sum_{i \neq j} |M_{ij}| + \lambda(1-\alpha) \sum_{k=1}^{K} \sum_{i \neq j} |H_{ij}^{(k)}| \\ \text{s.t.} \quad & \Theta^{(k)} = H^{(k)} + M, \quad \Theta^{(k)} \succeq 0. \end{aligned} \tag{4}$$

2.3. Augmented Lagrange solver

To solve problem (4), we could construct an augmented Lagrange function, and the objective function is transferred into the following optimization problem:

$$\begin{aligned} \text{minimize}_{\{\Theta^{(k)}, H^{(k)}, M\}} \quad & \sum_{k=1}^{K} n_k \left[\text{tr}\left(S^{(k)} \Theta^{(k)} \right) - \log\left\{ \det\left(\Theta^{(k)} \right) \right\} \right] \\ & + \lambda(1-\alpha) \sum_{k=1}^{K} \sum_{i \neq j} |H_{ij}^{(k)}| + \lambda\alpha K \sum_{i \neq j} |M_{ij}| \\ & + \sum_{k=1}^{K} \left\langle Y^{(k)}, \quad \Theta^{(k)} - \left(M + H^{(k)} \right) \right\rangle + \tfrac{\mu}{2} \left\| \Theta^{(k)} - H^{(k)} - M \right\|_F^2, \end{aligned} \tag{5}$$

where $\langle A, B \rangle = \text{tr}\left(AB^T \right), \|A\|_F$ represents Frobenius norm of matrix A, μ is the penalty factor, and $Y^{(k)}$ is Lagrange multiplier. Let $V^{(k)} = Y^{(k)}/\mu$, then the objective function of problem (5) takes the following form.

$$\begin{aligned} L\left(\left\{ \Theta^{(k)}, H^{(k)}, M, V^{(k)} \right\} \right) = & \sum_{k=1}^{K} n_k \left[\text{tr}\left(S^{(k)} \Theta^{(k)} \right) \right. \\ & - \log\left\{ \det\left(\Theta^{(k)} \right) \right\}] + \lambda\alpha K \sum_{i \neq j} |M_{ij}| + \lambda(1-\alpha) \sum_{k=1}^{K} \sum_{i \neq j} |H_{ij}^{(k)}| \\ & + \sum_{k=1}^{K} \left(\frac{\mu}{2} \|\Theta^{(k)} - H^{(k)} - M + V^{(k)}\|_F^2 - \frac{\mu}{2} \|V^{(k)}\|_F^2 \right). \end{aligned} \tag{6}$$

For $k = 1, \ldots, K$, minimizing $L\left(\left\{ \Theta^{(k)}, H^{(k)}, M, V^{(k)} \right\} \right)$ with other parameters fixed could update $\Theta^{(k)}, H^{(k)}, M, V^{(k)}$ in turn. The workflow is shown in Algorithm 1, where the subscript i represents the index of current iteration.

In order to accelerate the convergence of the algorithm, we let μ slightly increase as $\mu_{(i+1)} \longleftarrow 1.2\mu_{(i)}$. And the algorithm will stop when $\sum_k \|\Theta_{(i)}^{(k)} - \Theta_{(i-1)}^{(k)}\|_F^2 / \sum_k \|\Theta_{(i-1)}^{(k)}\|_F^2 < 10^{-5}$. We will discuss how to calculate $S^{(k)}$ for observed data with missing values in the following section.

Algorithm 1: The Augmented Lagrange Method for Solving JGCGM

```
input: $S^{(k)}, n_k, \mu, \lambda, \alpha$
initialization: $\mu = 0.1, \Theta_{(0)}^{(k)} = I, V_{(0)}^{(k)} = 0, H_{(0)}^{(k)} = 0, M = 0$, for
  $k = 1, 2, \ldots, K$, where sub-scripts $(i)$ presents the number of
  iteration in the algorithm.
1: whilethe algorithm is no converged do 2:
  $\left\{ \Theta_{(i+1)}^{(k)} \right\} \longleftarrow \arg\min_{\Theta} L_\mu \left( \left\{ \Theta^{(k)} \right\}, \left\{ H^{(k)} \right\}, M_{(i)}, \left\{ V_{(i)}^{(k)} \right\} \right)$.
3: $\left\{ H_{(i+1)}^{(k)} \right\} \longleftarrow \arg\min_{H} L_\mu \left( \left\{ \Theta_{(i+1)}^{(k)} \right\}, \left\{ H^{(k)} \right\}, M_{(i)}, \left\{ V_{(i)}^{(k)} \right\} \right)$.
4: $M_{(i+1)} \longleftarrow \arg\min_{M} L_\mu \left( \left\{ \Theta_{(i+1)}^{(k)} \right\}, \left\{ H_{(i+1)}^{(k)} \right\}, M, \left\{ V_{(i)}^{(k)} \right\} \right)$.
5: $\left\{ V_{(i+1)}^{(k)} \right\} \longleftarrow \left\{ V_{(i+1)}^{(k)} \right\} + \left\{ \Theta_{(i+1)}^{(k)} \right\} - \left\{ H_{(i+1)}^{(k)} \right\} - M_{(i+1)}$.
6: $i \longleftarrow i + 1$
7: end while
```

2.4. Estimation of correlation matrix from data with missing values

The input $S^{(k)}$ stand for the sample covariance matrices for Gaussian distributions and could be used directly for maximizing the likelihood function in Gaussian graphical models. But for nonparanormal distribution, $S^{(k)}$ can be regarded as correlation matrix. Generally, Kendall's tau coefficient [43] could estimate the approximate value of the correlation [41].

Computing Kendall's tau coefficient requires complete observations. However, scRNA-seq data contains a lot of missing values, which are also known as dropouts. In some scRNA-seq datasets, the dropout rate is even higher than 50% [44]. The missing values inherent in scRNA-seq data will inevitably bring large bias into the estimation of gene networks. To reduce the influence of missing values, we employ a modified Kendall's tau [37]. We firstly define a boolean coefficient $b_{ij}$to indicate whether a pair of values is valid, where $b_{ij} = 1$ if and only if the expression levels of genes i and j are observed. The modified Kendall's tau is calculated as follows.

$$\hat{\tau}_{jk} = \frac{1}{n_{jk}(n_{jk} - 1)} \sum_{i,i'=1, \; i \neq i'}^{n} b_{ij} b_{ik} b_{i'j} b_{i'k} \text{sign}\left(\left(x_i^j - x_{i'}^j \right) \left(x_i^k - x_{i'}^k \right) \right). \tag{7}$$

Here $i, i' = 1, \ldots n$ represent the indices of samples, $n_{jk} = \sum_{i=1}^{n} b_{ij} b_{ik}$ denotes the number of valid sample pairs for gene pair (j, k). Eq. (7) shows that calculate the modified Kendall's tau only utilizes the observed values, and the missing values are excluded.

Then the coefficient matrix $\widehat{S}^\tau = \left[\widehat{S}_{jk}^\tau \right]$ can be estimated as follows [43,45]:

$$\widehat{S}_{jk}^\tau = \begin{cases} \sin\left(\frac{\pi}{2} \hat{\tau}_{jk} \right) & \text{if} j \neq k \\ 1 & \text{if} j = k \end{cases} \tag{8}$$

where $\sin(\cdot)$ represents the sine function. Note that the estimated $\widehat{S}^\tau$ from Eqs. (7) and (8) may not be semi-positive definite due to the dropouts, and we could project it onto the cone of positive semi-definite matrix. That is, to solve the following optimization problem:

$$\widehat{S} = \mathrm{argmin}_{S \succcurlyeq 0} \left\| \widehat{S}^{\tau} - S \right\|_{\infty}. \tag{9}$$

$\widehat{S}$ in Eq. (9) are the final results of the estimated correlation matrices, which is the input of Algorithm 1.

3. Simulation studies

3.1. Data generation

3.1.1. Poisson graphical models for count data

In order to verify the effectiveness of our proposed model, we generate simulation data with known network structures. Here, following the work of [46,33], we assume that the discrete counts of scRNA-seq are multivariate Poisson distributed, and the network structure of the generated data satisfy Poisson graphical model [46]. Instead of the complete multivariate Poisson model, we use a degenerated version—the so-called multivariate Poisson model with two-way covariance structure [47] in our study. In the following of the paper, the model we used is roughly referred to as Poisson graphical model (PGM) for simplicity.

PGM models each value of a multivariate Poisson distributed vector as the sum of two Poisson distributed "source" variables. For example, suppose there are three independent Poisson source variables y_1, y_2 and y_3. The sum-up of variables $x_1 = y_1 + y_3$ and $x_2 = y_2 + y_3$ also satisfy Poisson distribution, and we could see that x_1, x_2 are dependent if $E(y_3) \neq 0$ because of the fact that $cov(y_1, y_2) \neq 0$. In other words, the combination of source variables encodes the underlying dependency of each pair of dimensions in a multivariate Poisson distributed vector.

Generally, a multivariate Poisson matrix X that contains n samples (observations) of p-dimensional vectors can be generated as $X = YB$, where Y is a $n \times [p + p(p-1)/2]$ matrix to simulate pairs of p source variables (contains $p(p-1)/2$ pairs and p original variables) and B defined how these variables are combined. By taking the form of: $B = \left[I_{(p)}, P_{erm} \odot \left(\mathbf{1}_{(p)} \mathrm{tri}(A)^T\right)\right]^T$, the underlying network of attributes in X is totally determined by the $p \times p$ adjacency matrix A. Here, $[\cdot,\cdot]$ represents to concatenate two matrices horizontally. P_{erm} is a $p \times p(p-1)/2$ permutation matrix; all its columns have 2 "1" s with others are "0" s. We illustrate a permutation matrix with $p = 4$ as follows:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Go back to the notation of B : $\odot$ stands for element-wise matrix product. $\mathbf{1}_{(p)}$ is an all "1" p-dimensional vector. tri(A) denotes the vector formed by the upper triangle element from A.

3.1.2. Data generation

Taking $K = 2$ cell types as examples. Note that in the following simulation studies, unless specifically mentioned, the number of nodes is set to $p = 100$. The default number of samples for each cell type is set to $n_k = 400$. When K, p, n_k is given, the datasets $X^{(k)}, (k = 1, 2)$ are generated according to the following steps:

1. **Generate the true network $A^{(1)}$ for the first cell type.** Since scale-free network is more similar to real biological networks, we first generate a scale-free network as the true network $A^{(1)}$. Barabasi-Albert algorithm in *R.igraph* is used to generate the random network with $p = 100$, with parameter $power = 0.01$.
2. **Generate the true network $A^{(2)}$ for the second cell type.** $A^{(2)}$ is obtained by randomly eliminating 10% of the edges from $A^{(1)}$.
3. **Generate count data $X^{(k)}$.** For $k = 1, 2, X^{(k)} = Y^{(k)} B^{(k)}$, where $Y_{ij}^{(k)} \overset{iid}{\sim} P_{oisson}(1), B^{(k)} = \left[I_{(p)}, P_{erm} \odot \left(\mathbf{1}_{(p)} \mathrm{tri}\left(A^{(k)}\right)^T\right)\right]^T$.

The impact of missing values on network inference is another focus of our study. For simplicity, we force the dropouts to occur only on small real counts value (less than 5). To simulate the dropout events in scRNA-seq, the read counts will be set to zero with a probability δ. We set 3 different values (i.e., $\delta = 0.1, 0.3, 0.5$) throughout the simulation to investigate the impact comprehensively.

In the following experiments, the results are obtained via 10 random generations of the data.

3.2. Compared methods and evaluation metrics

We compare our proposed JGCGM with the following 6 state-of-the-art network inference algorithms:

- Group Graphical Lasso (**GGL**) [48], a method for joint estimation of multiple Gaussian graphical models. We use Kendall's tau coefficient to estimate the correlation matrix for GGL since scRNA-seq data does not follow Gaussian distribution.
- Local Poisson Graphical Modal (**LPGM**) [49], a network inference method designed for single Poisson graphical model.
- Detecting Shared and Individual parts of MULtiple graphs Explicitly (**SIMULE**) [36], a method automatically infers both specific edge patterns and shared interaction preserved among all cell types.
- GEne Network Inference with Ensemble of trees (**GENIE3**) [50], winner of the DREAM5 network challenge, which is designed for bulk data but could also be applied on single cell sequencing.
- Gene Regulatory Network inference using gradient Boosting machine (**GRNBoost2**) [51], a scalable and more efficient alternative of GENIE3.
- Gene regulatory network inference uses Partial Information Decomposition and Context (**PIDC**) [52], a method for single cell transcriptomic data with heterogeneity information considered.

GGL, LPGM, SIMULE are implemented in R language, GENIE3 and GRNBoost2 are implemented by python package Arboreto [51] and PIDC is implemented in Julia.

We introduce three measurements to evaluate the accuracy of the estimated networks: true positive rate (TPR, also named as Recall), false positive rate (FPR) and Precision. Let θ_{ij} and $\hat{\theta}_{ij}$denote the true value and the estimated value of the precision matrix. TPR and FPR are computed as follows:

$$TPR = \frac{\sum_{k=1}^{K}\sum_{i<j} I\left\{\hat{\theta}_{ij}^{(k)} \neq 0 \quad \text{and} \quad \theta_{ij}^{(k)} \neq 0\right\}}{\sum_{k=1}^{K}\sum_{i<j} I\left\{\theta_{ij}^{(k)} \neq 0\right\}},$$

$$FPR = \frac{\sum_{k=1}^{K}\sum_{i<j} I\left\{\hat{\theta}_{ij}^{(k)} \neq 0 \quad \text{and} \quad \theta_{ij}^{(k)} = 0\right\}}{\sum_{k=1}^{K}\sum_{i<j} I\left\{\theta_{ij}^{(k)} = 0\right\}}, \tag{10}$$

$$Precision = \frac{\sum_{k=1}^{K}\sum_{i<j} I\left\{\hat{\theta}_{ij}^{(k)} \neq 0 \quad \text{and} \quad \theta_{ij}^{(k)} \neq 0\right\}}{\sum_{k=1}^{K}\sum_{i<j} I\left\{\hat{\theta}_{ij}^{(k)} \neq 0\right\}}.$$

For Gaussian graphical model-based models, the entry in precision matrices $\theta_{ij} = 0$ means $|\theta_{ij}| < 10^{-5}$. For other methods, these measurements are calculated through binary elements in the adjacency matrices.

3.3. Parameter settings

We first discuss the parameter settings. For GENIE3, GRNBoost2 and PIDC, we use the default parameters provided by softwares. These methods output complete graphs where the connectivity between each pair of genes are measured by correlation coefficient, which can be binarized by taking a threshold. By changing the value of the threshold, we can obtain different network structures, which could be used to plot Precision-Recall Curves (PRC) and Receiver Operating characteristic Curves (ROC). LPGM only has one parameter controlling the network sparsity. Fine-tuning this parameter is similar to changing the threshold in GENIE3, GRNBoost2 or PIDC.

Note that all the three Gaussian Graphical Models (GGM) based methods (i.e., JGCGM, GGL and SIMULE) have two tuning parameters, the first ones (λ_1 in JGCGM, ω_1 in GGL and λ in SIMULE) controls the sparsity of the estimated networks, and the second ones (α in JGCGM, ω_2 in GGL and ϵ in SIMULE) controls the heterogeneity among different cell types. Fine-tuning the sparsity parameters with the second parameters being fixed can control the sparsity of the estimated networks, which is similar to adjusting the thresholds in GENIE3, GRNBoost2 or PIDC. The value of ϵ in "SIMULE" is set to 0.5. For GGL, we set ω_2 to be 0.02 according to the experiments conducted in [26].

We analyze the effect of the proportion factor α in JGCGM. Fig. 2 shows the average results on synthetic data with $n_k = 400, p = 100$ and $\delta = 0.3$. Lines in different colors and markers shows the performances with different proportion factor. According to these experiment results, we find that when α is properly configured within a smaller range $[0.1, 0.35]$, JGCGM is not very sensitive to this parameter. Thus, for simplicity of the following simulation studies, α is roughly set to be 0.3.

3.4. Effects of data imputation

Here we conduct a simple experiment to test how imputation methods affect network inference. According to the study of Hou et al. [9], we choose 4 top imputation methods: SAVER [53], DrImpute [12], MAGIC [13] and McImpute [54]. We test the performance on three datasets, with $\delta = 0.1, 0.3, 0.5$ respectively. Each dataset is imputed with these 4 methods and 5 datasets are obtained (one is the original dataset without imputation). We infer the networks by applying 6 algorithms to these datasets, where JGCGM is not included in this test since JGCGM is designed for data with missing values.

The performances are shown by TPR-FPR curves in Fig. 3. McImpute dramatically improves the performance for all methods, followed by DrImpute which is also helpful for network inference in most situations. The only exception with PIDC for datasets with large dropout probability ($\delta = 0.5$ in Fig. 3 F(3)). These two methods fill in the "0" s in the original data while keeping other nonzero elements unchanged. The performances get worse if the data is imputed by SAVER or MAGIC. Specifically, SAVER slightly decreases the performance of network inference and MAGIC significantly decreases the performances of network inference. This may due to the build-in transformations in these algorithms destroy the intrinsic distribution of data.

Overall, McImpute achieves the best performance for network inference. Thus, we pick it as the default imputation method in the following experiments.

3.5. Comparison with other state-of-the-art methods

By calculating the average area under ROC (AUROC) and PRC (AUPRC), we could compare the accuracy of different network inference methods. T-tests are also conducted to evaluate the significance of the improvement of our JGCGM over other methods. Note that the true networks are so sparse that the number of negative samples is much larger than positive samples. In such cases, AUPRC is more informative and authoritative than AUROC. Thus, we mainly utilize AUPRC to measure the accuracy of various methods. Accordingly, we put the results in terms of AUROC into Supplementary Materials.

For Gaussian graphical-based models, log-transformation is also widely used to reduce the skewness of datasets. Thus, we also carry out experiments to test the effect of log-transformation on network inference. The results (Figure Supplementary S1) show that Pearson correlation with log-transformation is generally worse than Kendall's tau for network estimation. Therefore, we do not consider log-transformation in the following experiments.

3.5.1. Network inference for data with different sample sizes

Here, we focus on network inference of two cell subgroups. To study the impact of the sample size on network inference, we generate four datasets with different sample sizes, i.e., $n_k = 100, 400, 1000, 2000$ (for $k = 1, 2$). The number of genes is fixed at $p = 100$ and dropout probability δ is set to be 0.1, 0.3 and 0.5.

Tables 1 and 2 show the performance of various methods in terms of AUPRC. We can find from these tables that JGCGM dominates other methods in almost all cases except for $n_k = 1000, 2000$ and $\delta = 0.1$, where GGL performs slightly better than JGCGM. Nevertheless, the p-values indicate that the advantage of GGL is not so significant (0.515 and 0.334). Compared to other methods except for JGCGM, LPGM performs better when sample size is small. But the increase in the number of samples does not bring a significant improvement in its performance. For other 5 methods, i.e., GGL, SIMULE, GRNBoost2, GENIE3 and PIDC, none of them dominates others in all cases.

The performance of all methods is improved with the increases of sample size. All methods keep at an acceptable level when the dropout rate is low. However, when the dropout rate becomes higher, misleading information will break the original distribution of data to a certain degree, and reduce the accuracy of all methods. In the extreme situation when the sample size is small ($n_k = 100$) but the dropout rate is large ($\delta = 0.5$), GRNBoost2 and PIDC is only slightly better than "randomly guess" (AUROC = 0.509 and 0.549 respectively in Supplementary Table S1).

Consistent with previous tests in Section 3.4, all methods except JGCGM have been improved after data imputation. The results suggest that appropriate imputation is beneficial for network inference. However, even with the help of imputation, other methods still fall behind JGCGM. Note that GGL and SIMULE achieve pretty good performance when the dropout rate is low, but suffer a sharp decline in accuracy if the dropout rate become larger (e.x. AUPRC from 0.657 to 0.109 with δ changing from 0.1 to 0.3 for GGL). As a comparison, the proposed JGCGM, which is also a Gaussian graphical model-based method, is not affected by the dropout rate.

3.5.2. Network inference for data with outliers

We add some outliers in data to test the robustness of the algorithms. Specifically, every count value will be added a constant value $c_o$with a probability of 5%. We set $c_o$to be 2/3 of the maximum count value in the corresponding cell type. Here we set the sample size to be $n_k = 400$.

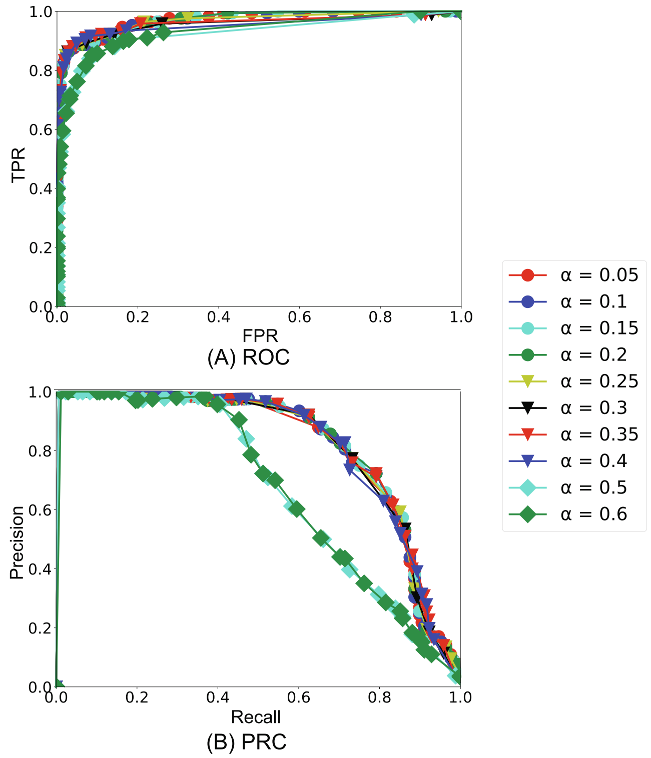

Fig. 2. The effect of different tuning parameters (proportion factors α) in JGCGM on estimation performance. (The simulated data are generated with $K = 2, n_k = 400$ and $\delta = 0.3$).

Fig. 3 shows the average prediction results of 7 algorithms. We can see that all algorithms are disturbed by the outliers, and the accuracy suffers severely reduction. Especially when the sample size is small and the dropout rate is high, no algorithm can achieve satisfactory results. Even so, JGCGM is still better than other algorithms except for one dataset with $\delta = 0.1$.

Imputation method still works for data with outliers. Different methods benefit from McImpute distinctly when outliers exist in the data.

3.5.3. Network inference on multiple cell subgroups

We then conduct experiments on data with 3 and 4 cell subgroups. We set the sample size to be $n_k = 400$, and the number of genes to be $p = 100$. The procedure of generating synthetic data is similar to above Section 3.1.2, apart from the little difference on network construction: We first construct a sparse scale-free network as the basic network structure for all cell types $A^{(0)}$, which contains p nodes and a total of n_{edge}edges. Then we randomly eliminate $0.1n_{edge}$edges from the basic network. For each time we eliminate edges from $A^{(0)}$, and generate a network for a new cell type $A^{(k)}$ (for $k = 1, 2, 3, 4$). By repeating this step, we can obtain 3 or 4 networks and generate the corresponding read counts data accordingly. We do not impute the data in these experiments, as its impact has been discussed before.

Table 3 shows the prediction results of various algorithms, with the number of cell types to be 3 and 4 respectively. Dropout events still bring a lot of trouble to the network inference. JGCGM is relatively less affected by it, and achieves the best prediction accuracy in most cases. Among all the compared methods, GENIE3 achieves the best results, followed by SIMULE, GGL and GRNBoost2. LPGM and PIDC perform the worst according to AUPRC, whereas LPGM gets quite higher AUROC (See Table S4).

According to these experiment results, we find that only JGCGM can get better performance when the number of cell types grows. These results demonstrate that JGCGM is able to make use of the similarities between different cell types to improve the accuracy of network inference.

4. Real data analysis

4.1. Parameter selection

For real-world data analysis, the hyperparameters in JGCGM (λ, α) should be chosen carefully. According to the discussion in Section 3.3, the search range of α is set to $[0.1, 0.35]$ and the optimal α^* is chosen via minimizing Akaike information criterion (AIC):

$$\mathrm{AIC} = \sum_{k=1}^{K}\left\{n_k \mathrm{tr}\left(\widehat{S}^{(k)}\hat{\Theta}^{(k)}\right) - n_k \log\det\left(\hat{\Theta}^{(k)}\right) + 2n_e^{(k)}\right\},$$

where K is the number of cell types, $n_e^{(k)}$ is the edge number in the estimated graph, $\widehat{S}^{(k)}$ is the estimated correlation matrix, and $2n_e^{(k)}$ equals the number of non-zero entities of the estimated precision matrix $\hat{\Theta}^{(k)}$.

When we have got α^*, we use the stability selection method [55] to select parameters λ. Specifically, at each time, 50% of the samples are randomly selected from each group to form a sub-dataset. This operation repeats C times with each time sampling a sub-dataset. All C sub-datasets are analyzed with fixed parameters (λ, α^*) and then obtain networks $\left\{\hat{\Theta}_s^{(k)}(\lambda, \alpha^*)\right\}$. Let $\overline{B}_{ij}^k(\lambda) = (1/C)\sum_{s=1}^{S}\mathbf{1}\left(\hat{\Theta}_{s,ij}^k(\lambda, \alpha_{\mathrm{opt}}) \neq 0\right)$, we could measure the stability score of these parameters Stab by computing:

$$\mathrm{Stab}(\lambda) = \sum_{k=1}^{K}\left(\sum_{i<j}\overline{B}_{ij}^k(\lambda)\left(1 - \overline{B}_{ij}^k(\lambda)\right)/\binom{p}{2}\right) \tag{11}$$

We first choose a set of candidate values for λ, then compute the stability scores for all these candidate values and finally get the optimal λ^* through the following equation:

$$\lambda^* = \arg\min_{\gamma\in\Lambda}\left\{\max_{\lambda\geqslant\gamma}\mathrm{Stab}(\lambda) \leqslant \beta\right\}. \tag{12}$$

For all real data sets, we set $C = 20, \beta = 0.1$ in the procedure of stability selection.

4.2. Differential network between H1-hESC and NPC

Joint network inference methods can be well suited for differential network analysis. We study the differential network between H1 human embryonic stem cells (H1-hESC) and neural progenitor cells (NPC). The dataset that contains 212 H1-hESCs and 173 NPCs originated from the literature [56], which can be downloaded from GEO database [57] (with ID GSE75748). Since the NPCs in this dataset are differentiated from H1-hESCs, we focus on a subset of 20 genes that are involved in human embryo development [20]. Another 70 type-specific genes and 44 “dark” genes, which have a significant difference between and control samples not in network degree level [20], are merged into the subset. By filtering out the duplicated and unexpressed genes from the subset, we finally get a list of 113 genes for our study.

The optimal parameters $(\alpha^* = 0.35, \lambda^* = 0.3)$ are chosen via AIC and stability selection. Fig. 4 presents the differential network between H1-hESC and NPC predicted by JGCGM. 6 genes (CST1, DLK1, PHC1, TDGF1, VSNL1 and L1TD1) with the highest degrees (> 10) are marked with yellow colors. Among these genes, 5 of them (except for CST1) are found to be cell type-specific [20].

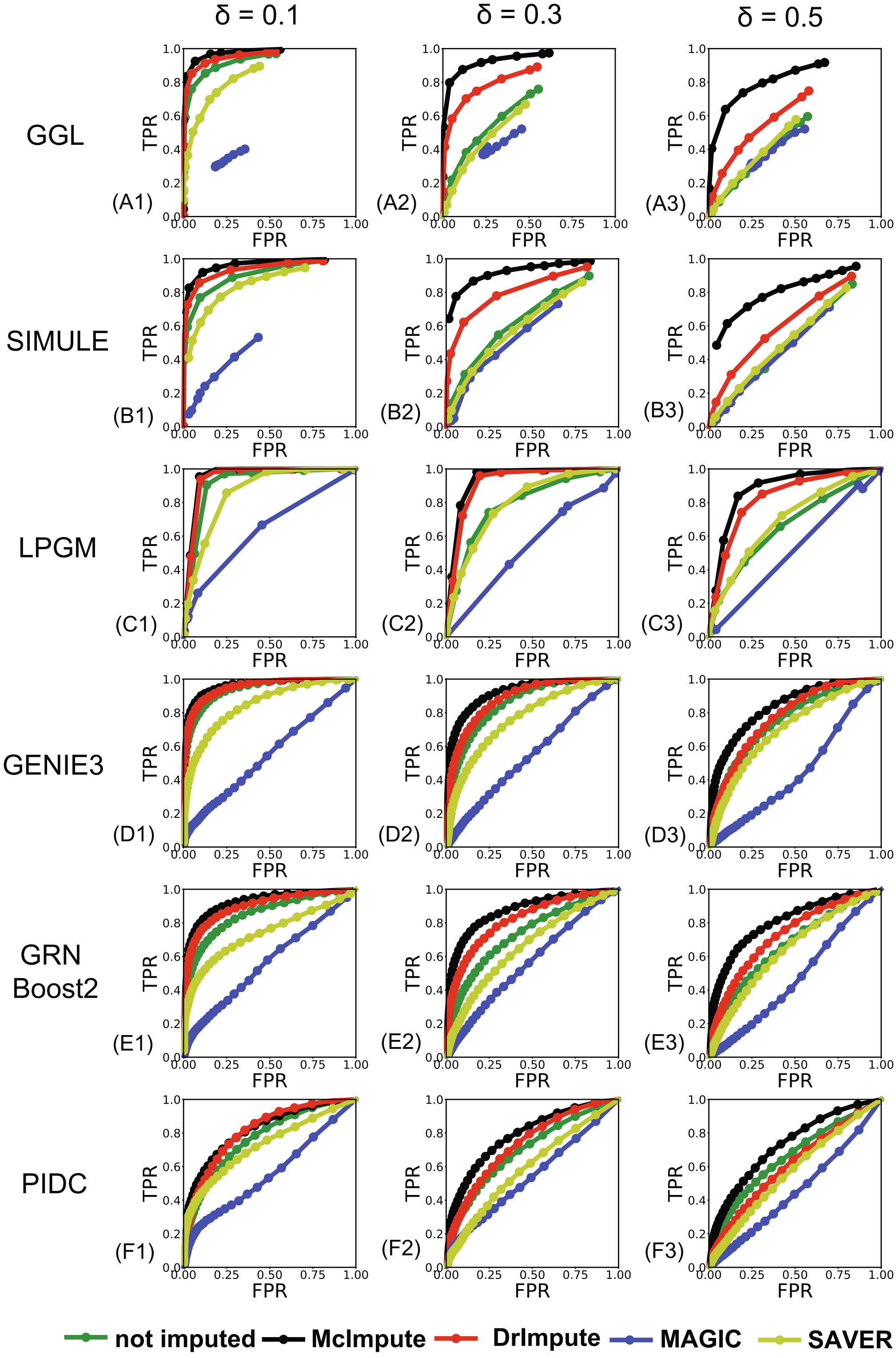

Fig. 3. Comparison of the imputation methods on network inference (Synthetic data with $n_k = 400$ and $p = 100$).

Besides, JGCGM is able to discover the two dark genes for H1-hESC (TDGF1 and VSNL1) defined in the study of [20]. Although the differential network does not present the significance of four dark genes for NPC (TMEM97, SGPL1, ICAM1, ARSA), all these genes are with higher rank of degrees in the NPC-specific network than in the H1-hESC network (see Table S5), which also validate the effectiveness of JGCGM.

4.3. Mouse embryonic stem cell differentiation

Cellular differentiation decisions are controlled by complex regulatory interactions, and understanding regulatory mechanism remains to be a major challenge. Here we study the case of embryonic stem cells differentiation by single cell RNA sequencing data from mouse. The dataset originated from the literature [58], which

Table 1
AUPRC of various methods on synthetic data with small sample size ($n_k \geqslant 1000, p = 100$).

n_k	δ	Imputation	Value	JGCGM	LPGM	GGL	SIMULE	GRNBoost2	GENIE3	PIDC
1000	0.1	No	Mean	**0.941**	0.233	0.896	0.825	0.391	0.855	0.520
			Std	0.020	0.027	0.023	0.028	0.015	0.009	0.021
			p-value	–	1.11E−22	2.66E−04	7.68E−09	5.31E−23	5.89E−10	1.02E−19
		Yes	Mean	–	0.312	**0.946**	0.920	0.849	0.899	0.624
			Std	–	0.020	0.010	0.013	0.022	0.009	0.020
			pvalue	–	5.56E−23	5.15E−01	1.48E−02	2.21E−08	1.66E−05	8.50E−18
	0.3	No	Mean	**0.924**	0.172	0.299	0.259	0.188	0.613	0.383
			Std	0.013	0.022	0.024	0.031	0.011	0.014	0.021
			p-value	–	3.26E−25	4.61E−23	4.19E−22	4.58E−28	1.29E−20	1.06E−22
		Yes	Mean	–	0.299	0.787	0.753	0.662	0.766	0.512
			Std	–	0.030	0.026	0.030	0.024	0.015	0.022
			p-value	–	7.17E−22	3.45E−11	5.15E−12	2.25E−16	5.58E−15	2.20E−20
	0.5	No	Mean	**0.908**	0.127	0.019	0.045	0.098	0.296	0.294
			Std	0.018	0.031	0.005	0.005	0.006	0.013	0.026
			p-value	–	7.15E−23	7.35E−29	1.36E−28	5.46E−28	1.46E−24	7.75E−22
		Yes	Mean	–	0.184	0.378	0.340	0.283	0.390	0.191
			Std	–	0.040	0.043	0.043	0.030	0.030	0.027
			p-value	–	1.00E−20	9.94E−18	2.16E−18	2.57E−21	8.83E−20	5.27E−23
2000	0.1	No	Mean	0.945	0.244	**0.965**	0.790	0.863	0.931	0.786
			Std	0.022	0.030	0.013	0.066	0.010	0.006	0.015
			p-value	–	9.56E−22	3.10E−02	2.91E−06	4.59E−09	6.75E−02	4.94E−13
		Yes	Mean	–	0.214	**0.959**	0.840	0.927	0.946	0.804
			Std	–	0.069	0.035	0.142	0.011	0.006	0.015
			pvalue	–	6.82E−17	3.34E−01	4.09E−02	3.61E−02	8.84E−01	4.21E−12
	0.3	No	Mean	**0.949**	0.169	0.481	0.420	0.481	0.806	0.653
			Std	0.013	0.024	0.032	0.031	0.013	0.009	0.020
			p-value	–	4.50E−25	4.33E−19	1.91E−20	3.16E−24	4.05E−16	1.74E−18
		Yes	Mean	–	0.223	0.795	0.661	0.712	0.852	0.566
			Std	–	0.024	0.028	0.025	0.022	0.010	0.028
			p-value	–	2.35E−24	1.09E−11	4.32E−17	2.86E−16	5.45E−13	2.12E−18
	0.5	No	Mean	**0.961**	0.132	0.020	0.053	0.196	0.513	0.557
			Std	0.010	0.039	0.006	0.016	0.020	0.015	0.017
			p-value	–	1.67E−22	3.32E−33	3.56E−29	1.82E−26	5.89E−24	1.60E−22
		Yes	Mean	–	0.159	0.273	0.250	0.284	0.492	0.307
			Std	–	0.032	0.028	0.025	0.018	0.015	0.019
			p-value	–	1.37E−23	2.26E−23	2.20E−24	4.57E−26	3.48E−24	2.05E−25

can be downloaded from GEO database [57] (with ID GSE65525). In the original biological experiments, cells were sampled from 4 states, that is before leukemia inhibitory factor (LIF) withdrawal and after the withdrawal LIF for 2, 4, 7 days respectively. Since LIF is able to maintain cell pluripotency and inhibiting cell differentiation, it can be considered that the cells begin to gradually differentiate after LIF withdrawal.

According to the days of cell differentiation: 0 (before LIF withdrawal), 2, 4 and 7 the samples can be naturally divided into 4 groups, with each subgroup containing 933, 303, 683 and 798 cell samples respectively. We focus on 84 genes stem cell markers for the research [59,60]. Count values of the corresponding genes are extracted from the raw data in each cell group. Therefore we get four read count matrices act as the input of JGCGM for network inference subsequently.

We choose the parameters ($\alpha^* = 0.3, \lambda^* = 0.12$) for JGCGM via AIC and stability selection. Table 4 shows the nodes with the highest degree in the inferred networks at the optimal parameters. These high-ranking hub genes in the gene networks are of significant importance for maintaining cell life functions, which may be the potential key genes that affect cell differentiation. To seek from the candidates that change most during differentiation, we define a "changing score" of a gene as the difference between the maximum value and the minimum value of its ranking across cell subgroups. We set the threshold of changing score as $D_d \geqslant 5$, and eventually pick out 10 genes: Pou5f1 (5 for changing score), Fn1 (6), Gcm1 (9), Podx1 (5), Cd9 (6), Zfp42 (8), Utf1 (12), Lama1 (6), Gcg (49), Ifitm1 (6).

Most of the chosen genes are indeed related to cell differentiation: The Oct-4 protein encoded by the Pou5f1 gene plays an important role in the self-renewal of undifferentiated embryonic stem cells. Its over-expression and under-expression will result in cell differentiation, which is often regarded as a marker of undifferentiated cells [61]. FN1 plays an important role in tissue development and regeneration. Wang et al. [62] found that the fat pad stem cells' cartilage differentiation and fat differentiation ability decreased significantly after knocking out FN1. Gcm1 transcription factor covering FGF signal can promote the terminal differentiation of trophoblast stem cells [63], in addition, the gene can guide the change of cell fate by activating the glial development program in the multipotent precursor cells of the nervous system [64]. Zhang [65] experimented with mouse hematopoietic stem cells and found the surface expression of Podx1 can divide Flk1 overexpressing cells into different populations in the embryonic body that is being differentiated. The stromal cells expressing CD9 affect the physical interaction with hematopoietic cells and may be a factor that determines the degree of stem cell differentiation [66]. Zfp42 is reported as an

Table 2
AUPRC of various methods on synthetic data with outliers ($n_k = 400, p = 100$).

δ	Outlier	Imputation	Value	JGCGM	LPGM	GGL	SIMULE	GRNBoost2	GENIE3	PIDC
0.1	Without Outlier	No	Mean	**0.881**	0.210	0.657	0.570	0.354	0.597	0.178
			Std	0.009	0.023	0.023	0.016	0.022	0.020	0.019
			p-value	–	1.29E−24	5.04E−16	7.13E−21	4.79E−23	1.04E−18	3.47E−26
		Yes	Mean	–	0.286	0.815	0.762	0.546	0.710	0.282
			Std	–	0.016	0.020	0.019	0.026	0.020	0.031
			p-value	–	7.87E−26	4.91E−08	1.12E−12	3.17E−18	6.26E−15	1.54E−21
	With Outlier	No	Mean	0.276	0.089	**0.341**	0.232	0.050	0.058	0.106
			Std	0.027	0.010	0.026	0.013	0.003	0.002	0.009
			p-value	–	1.14E−13	4.68E−05	3.10E−04	1.53E−15	2.73E−15	4.55E−13
		Yes	Mean	–	0.085	**0.496**	0.302	0.055	0.061	0.145
			Std	–	0.008	0.024	0.048	0.003	0.003	0.012
			p-value	–	5.91E−14	3.38E−13	1.69E−01	2.42E−15	3.66E−15	7.71E−11
0.3	Without Outlier	No	Mean	**0.813**	0.134	0.109	0.085	0.138	0.278	0.124
			Std	0.017	0.019	0.020	0.012	0.013	0.015	0.012
			p-value	–	2.24E−24	1.98E−24	1.32E−26	8.59E−26	1.49E−23	4.86E−26
		Yes	Mean	–	0.271	0.706	0.097	0.393	0.527	0.206
			Std	–	0.019	0.026	0.011	0.023	0.026	0.023
			p-value	–	1.20E−22	4.61E−09	1.24E−26	9.60E−20	3.93E−16	1.13E−22
	With Outlier	No	Mean	**0.228**	0.083	0.048	0.055	0.043	0.047	0.072
			Std	0.029	0.010	0.012	0.007	0.003	0.002	0.005
			p-value	–	3.13E−11	1.35E−12	1.00E−12	2.22E−13	3.31E−13	4.60E−12
		Yes	Mean	–	0.086	**0.318**	0.186	0.049	0.054	0.100
			Std	–	0.007	0.024	0.041	0.004	0.003	0.006
			p-value	–	3.22E−11	1.45E−06	2.16E−02	4.46E−13	6.45E−13	1.35E−10
0.5	Without Outlier	No	Mean	**0.699**	0.093	0.026	0.035	0.081	0.133	0.096
			Std	0.040	0.010	0.005	0.004	0.005	0.010	0.010
			p-value	–	9.21E−20	9.26E−21	1.11E−20	4.28E−20	3.17E−19	9.54E−20
		Yes	Mean	–	0.184	0.384	0.060	0.226	0.274	0.132
			Std	–	0.017	0.028	0.007	0.022	0.023	0.010
			p-value	–	4.01E−18	1.76E−13	2.63E−20	4.13E−17	3.63E−16	3.07E−19
	With Outlier	No	Mean	**0.142**	0.070	0.026	0.036	0.038	0.042	0.066
			Std	0.014	0.007	0.004	0.003	0.002	0.001	0.004
			p-value	–	5.13E−11	4.16E−15	1.69E−14	1.75E−14	3.13E−14	6.42E−12
		Yes	Mean	–	0.086	0.125	0.101	0.049	0.048	0.069
			Std	–	0.013	0.017	0.054	0.003	0.001	0.006
			p-value	–	4.84E−08	2.95E−02	3.79E−02	1.40E−13	8.50E−14	2.54E−11

undifferentiated state marker with pluripotent hematopoietic stem cells [67]. Van et al. [68] found that knocking down UTF1 caused substantial delay or differentiation of embryonic stem cells and cancer cells. Only for three genes Lama1, Gcg, and Ifitm, we have yet to know their association with cell differentiation, wherein the excessive degree of Gcg is likely to be systematic bias.

The above case studies indicate that JGCGM is capable to reveal potential markers of cell differentiation, which verifies the effectiveness of our model in network inference.

5. Conclusion

With the development of single-cell RNA sequencing technologies, a large amount of single-cell RNA sequencing data become available, which enables the estimation of gene networks at single cell level. Due to the characteristics of scRNA-seq data, such as cellular heterogeneity and high sparsity caused by dropouts, joint estimation of multiple gene networks from scRNA-seq data remains a challenging task. Although Gaussian graphical model-based approaches have been widely used to infer gene networks, we are faced with two main issues when inferring gene networks from scRNA-seq data. The first one is how to handle the cellular heterogeneity. The other problem is the large proportion of dropouts (missing data) in scRNA-seq data.

To tackle these problems, we propose a new joint Gaussian copula graphical model (JGCGM) to jointly estimate multiple gene networks for multiple cell subgroups from scRNA-seq data. Our proposed model decomposes the gene network of each cell subgroup into two parts: common network and subgroup-specific network, which represent the heterogeneity and the homogeneity among different cell subgroups respectively. We use modified Kendall's tau to fulfill the estimation, which could make full use of the useful information from scRNA-seq data and keep away from the misleading information introduced by dropouts.

We compare proposed JGCGM with other methods on synthetic datasets. Our JGCGM outperforms other methods in most cases, which indicates the effectiveness of our proposed decomposition model on scRNA-seq data. Moreover, by using modified Kendall's tau, our JGCGM dominates other compared methods in most cases, which demonstrate its ability in handling dropout events. The impact of data imputation has also been studied in simulation studies. Among the selected imputation approach, McImpute dominates others in all circumstance and MAGIC is detrimental to network inference. Even though imputation could enhance the accuracy of network inference, its performance still falls behind from the proposed JGCGM. These results demonstrate the effectiveness of our model on scRNA-seq data.

We also predict the differential network between H1 human embryonic stem cells and neural progenitor cells, and study cell differentiation from mouse embryonic stem cells. The results of

Table 3
AUPRC of various methods on synthetic data with multiple cell types ($n_k = 400, p = 100$).

δ	K	Imputation	Value	JGCGM	LPGM	GGL	SIMULE	GRNBoost2	GENIE3	PIDC
0.1	3	No	Mean	**0.592**	0.214	0.443	0.524	0.343	0.581	0.165
			Std	0.030	0.013	0.033	0.034	0.018	0.011	0.012
			p-value	–	5.01E−18	7.14E−09	2.65E−04	2.46E−14	3.11E−01	4.69E−19
		Yes	Mean	–	0.276	0.602	**0.692**	0.491	0.656	0.241
			Std	–	0.018	0.032	0.022	0.020	0.012	0.015
			p-value	–	4.60E−16	4.97E−01	0.000	0.000	0.000	0.000
	4	No	Mean	**0.602**	0.220	0.338	0.513	0.334	0.577	0.167
			Std	0.010	0.017	0.027	0.026	0.015	0.012	0.009
			p-value	–	6.47E−22	2.72E−16	1.89E−08	9.29E−20	1.16E−04	4.99E−26
		Yes	Mean	–	0.270	0.495	0.652	0.467	0.642	0.233
			Std	–	0.018	0.023	0.033	0.018	0.009	0.011
			p-value	–	1.90E−20	1.41E−10	3.97E−04	1.11E−13	4.01E−08	4.66E−24
0.3	3	No	Mean	**0.502**	0.129	0.059	0.084	0.142	0.276	0.113
			Std	0.022	0.012	0.022	0.020	0.012	0.017	0.010
			p-value	–	7.17E−20	1.63E−19	1.88E−19	1.55E−19	2.91E−15	2.04E−20
		Yes	Mean	–	0.227	0.373	0.433	0.280	0.401	0.145
			Std	–	0.010	0.026	0.028	0.021	0.025	0.018
			p-value	–	8.88E−18	1.49E−09	2.02E−05	2.12E−14	4.66E−08	1.39E−18
	4	No	Mean	**0.542**	0.141	0.046	0.090	0.137	0.271	0.114
			Std	0.031	0.013	0.012	0.015	0.010	0.013	0.009
			p-value	–	2.75E−18	5.66E−20	5.37E−19	1.46E−18	2.63E−15	4.09E−19
		Yes	Mean	–	0.217	0.265	0.377	0.255	0.368	0.127
			Std	–	0.013	0.041	0.035	0.012	0.018	0.011
			p-value	–	1.29E−16	3.11E−12	3.65E−09	8.78E−16	1.73E−11	1.09E−18
0.5	3	No	Mean	**0.321**	0.102	0.024	0.045	0.077	0.125	0.095
			Std	0.036	0.018	0.004	0.003	0.004	0.008	0.005
			p-value	–	4.08E−12	3.25E−15	1.09E−14	9.67E−14	5.45E−12	3.81E−13
		Yes	Mean	–	0.149	0.128	0.153	0.136	0.174	0.079
			Std	–	0.014	0.021	0.027	0.005	0.012	0.004
			p-value	–	1.12E−10	5.98E−11	1.61E−09	1.18E−11	9.26E−10	1.14E−13
	4	No	Mean	**0.415**	0.102	0.017	0.043	0.078	0.129	0.090
			Std	0.037	0.021	0.004	0.002	0.004	0.004	0.007
			p-value	–	1.54E−14	2.20E−17	6.68E−17	4.16E−16	7.20E−15	9.49E−16
		Yes	Mean	–	0.153	0.085	0.124	0.123	0.162	0.070
			Std	–	0.013	0.019	0.019	0.007	0.008	0.004
			p-value	–	8.16E−14	3.91E−15	3.71E−14	6.16E−15	8.21E−14	2.69E−16

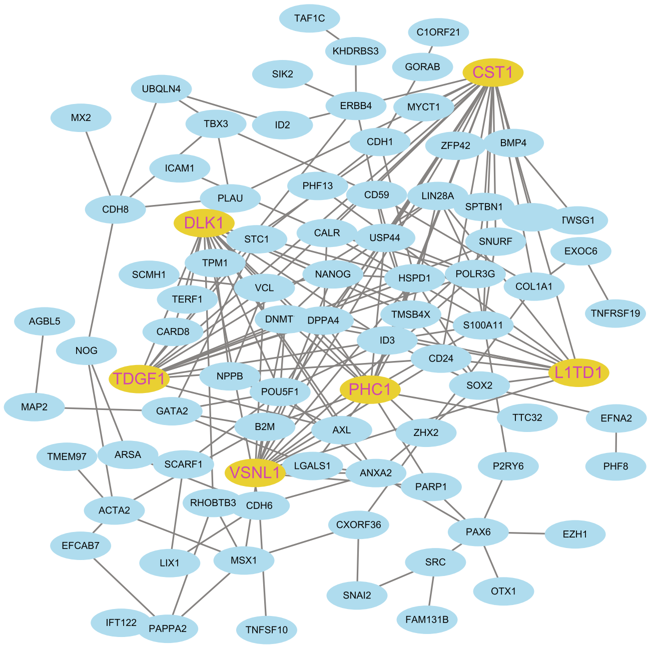

Fig. 4. The differential network between H1 human embryonic stem cells (H1-hESC) and neural progenitor cells (NPC) estimated by JGCGM. Hub genes with degrees greater than 10 are marked with yellow colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Top 15 nodes with the highest degree in 4 cell groups.

Rank	Day 0	Day 2	Day 4	Day 7
1	Lifr	Lifr	Lifr	Lifr
2	Pou5f1	Lamc1	Pou5f1	Lamc1
3	Pten	Pou5f1	Lamc1	Cd9
4	Fn1	Pten	Podxl	Fn1
5	Sox2	Sox2	Pten	Pten
6	Lamc1	Cd9	Sox2	Sox2
7	Gcm1	Fn1	Lama1	Pou5f1
8	Podxl	Lama1	Cd9	Lama1
9	Cd9	Podxl	Ifitm1	Podxl
10	Dnmt3b	Dnmt3b	Fn1	Dnmt3b
11	Zfp42	Gcm1	Dnmt3b	Gcm1
12	Utf1	Ifitm1	Nodal	Sox17
13	Lama1	Sox17	Foxa2	Ifitm1
14	Gcg	Zfp42	Sox17	Zfp42
15	Ifitm1	Nodal	Utf1	Diap2

the network prediction are mostly consistent with the known biological knowledge, which further confirms the effectiveness of our model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China [61602309, 61871272], Guangdong Basic and Applied Basic Research Foundation [2019A1515011384], Shenzhen Fundamental Research Program [JCYJ20170817095210760, JCYJ20190808173617147].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2020.09.004.

References

[1] Saeed MT, Ahmad J, Baumbach J, Pauling J, Shafi A, Paracha RZ, Hayat A, Ali A. Parameter estimation of qualitative biological regulatory networks on high performance computing hardware. BMC Syst Biol 2018;12(1):1–15. https://doi.org/10.1186/s12918-018-0670-y.

[2] Kim J. Validation and selection of ode models for gene regulatory networks. Chemometr Intell Lab Syst 2016;157:104–10. https://doi.org/10.1016/j.chemolab.2016.06.016.

[3] Sanchez-Castillo M, Blanco D, Tienda-Luna IM, Carrion M, Huang Y. A bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. Bioinformatics 2018;34(6):964–70. https://doi.org/10.1093/bioinformatics/btx605.

[4] Zhao H, Duan Z-H. Cancer genetic network inference using gaussian graphical models. Bioinf Biol Insights 2019;13. https://doi.org/10.1177/1177932219839402. 1177932219839402.

[5] Zhou Y, Qureshi R, Sacan A. Data simulation and regulatory network reconstruction from time-series microarray data using stepwise multiple linear regression. Network Model Anal Health Inf Bioinf 2012;1(1–2):3–17. https://doi.org/10.1007/s13721-012-0008-4.

[6] Liang K-C, Wang X. Gene regulatory network reconstruction using conditional mutual information. EURASIP J Bioinf Syst Biol 2008;2008(1):. https://doi.org/10.1155/2008/253894253894.

[7] Fu Y, Zhang F, Zhang X, Yin J, Du M, Jiang M, Liu L, Li J, Huang Y, Wang J. High-throughput single-cell whole-genome amplification through centrifugal emulsification and emda. Commun Biol 2019;2(1):1–10.

[8] Blencowe M, Arneson D, Ding J, Chen Y-W, Saleem Z, Yang X. Network modeling of single-cell omics data: challenges, opportunities, and progresses. Emerg Top Life Sci 2019;3(4):379–98. https://doi.org/10.1042/ETLS20180176.

[9] Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell rna-sequencing imputation methods. bioRxivdoi:https://doi.org/10.1101/2020.01.29.925974..

[10] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16(3):133–45. https://doi.org/10.1038/nrg3833.

[11] Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell rna-seq data. Nat Commun 2018;9(1):1–9. https://doi.org/10.1038/s41467-018-03405-7.

[12] Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. Drimpute: imputing dropout events in single cell rna sequencing data. BMC Bioinf 2018;19(1):220. https://doi.org/10.1186/s12859-018-2226-y.

[13] Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;174(3):716–29. https://doi.org/10.1016/j.cell.2018.05.061.

[14] Andrews TS, Hemberg M. False signals induced by single-cell imputation, F1000Research 7. doi:10.12688/f1000research.16613.1..

[15] Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinf 2018;19(1):232.

[16] Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods 2020;17(2):147–54.

[17] Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. Eleven grand challenges in single-cell data science. Genome Biol 2020;21(1):1–35. https://doi.org/10.1186/s13059-020-1926-6.

[18] Wang Y, Wu H, Yu T. Differential gene network analysis from single cell rna-seq. J Genet Genom=Yi chuan xue bao 2017;44(6):331. https://doi.org/10.1016/j.jgg.2017.03.001.

[19] Chiu Y-C, Hsiao T-H, Wang L-J, Chen Y, Shao Y-HJ. scdnet: a computational tool for single-cell differential network analysis. BMC Syst Biol 2018;12(8):124. https://doi.org/10.1186/s12918-018-0652-0.

[20] Dai H, Li L, Zeng T, Chen L. Cell-specific network constructed by single-cell rna sequencing data. Nucleic Acids Res 2019;47(11). https://doi.org/10.1093/nar/gkz172. e62–e62.

[21] Castro DM, De Veaux NR, Miraldi ER, Bonneau R. Multi-study inference of regulatory networks for more accurate models of gene regulation. PLoS Comput Biol 2019;15(1):. https://doi.org/10.1371/journal.pcbi.1006591e1006591.

[22] Jackson CA, Castro DM, Saldi G-A, Bonneau R, Gresham D. Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments. Elife 2020;9:. https://doi.org/10.1101/581678e51254.

[23] Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika 2007;94(1):19–35. https://doi.org/10.1093/biomet/asm018.

[24] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9(3):432–41. https://doi.org/10.1093/biostatistics/kxm045.

[25] Cai T, Liu W, Luo X. A constrained l1 minimization approach to sparse precision matrix estimation. J Am Stat Assoc 2011;106(494):594–607. https://doi.org/10.1198/jasa.2011.tm10155.

[26] Mohan K, London P, Fazel M, Witten D, Lee S-I. Node-based learning of multiple gaussian graphical models. J Mach Learn Res 2014;15(1):445–88. https://doi.org/10.1142/S0218194014500065.

[27] Ma J, Michailidis G. Joint structural estimation of multiple graphical models. J Mach Learn Res 2016;17(1):5777–824. https://doi.org/10.1093/biomet/asq060.

[28] Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika 2011;98(1):1–15. https://doi.org/10.1093/biomet/asq060.

[29] Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J Roy Stat Soc Ser B (Stat Methodol) 2014;76(2):373–97. https://doi.org/10.1111/rssb.12033.

[30] Zhang X-F, Ou-Yang L, Yan T, Hu XT, Yan H. A joint graphical model for inferring gene networks across multiple subpopulations and data types. IEEE Trans Cybern Doi: 10.1109/TCYB.2019.2952711..

[31] Church BV, Williams HT, Mar JC. Investigating skewness to understand gene expression heterogeneity in large patient cohorts. BMC Bioinf 2019;20(24):1–14. https://doi.org/10.1186/s12859-019-3252-0.

[32] Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.

[33] Gallopin M, Rau A, Jaffrézic F. A hierarchical poisson log-normal model for network inference from rna sequencing data. PloS One 8 (10). doi:10.1371/journal.pone.0077503..

[34] Liu H, Lafferty J, Wasserman L. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. J Mach Learn Res 2009;10(Oct):2295–328. https://doi.org/10.1145/1577069.1755863.

[35] Xue L, Zou H, et al. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Ann Stat 2012;40(5):2541–71. https://doi.org/10.1214/12-AOS1041.

[36] Wang B, Singh R, Qi Y. A constrained l1 minimization approach for estimating multiple sparse gaussian or nonparanormal graphical models. Mach Learn 2017;106(9–10):1381–417. https://doi.org/10.1007/s10994-017-5635-7.

[37] Wang H, Fazayeli F, Chatterjee S, Banerjee A. Gaussian copula precision estimation with missing values. Artif Intell Stat 2014:978–86.

[38] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT press 2009. https://doi.org/10.1002/9781118763117.ch9.

[39] Maathuis M, Drton M, Lauritzen S, Wainwright M. Handbook of graphical models. CRC Press; 2018.

[40] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. CRC Press 2015. https://doi.org/10.1111/insr.12167.

[41] Liu H, Han F, Yuan M, Lafferty J, Wasserman L, et al. High-dimensional semiparametric gaussian copula graphical models. Ann Stat 2012;40(4):2293–326. https://doi.org/10.1214/12-AOS1037.

[42] Lafferty J, Liu H, Wasserman L, et al. Sparse nonparametric graphical models. Stat Sci 2012;27(4):519–37. https://doi.org/10.1214/12-STS391.

[43] Kruskal WH. Ordinal measures of association. J Am Stat Assoc 1958;53(284):814–61. https://doi.org/10.2307/2281954.

[44] Qiu P. Embracing the dropouts in single-cell rna-seq data. bioRxiv 2018:. https://doi.org/10.1038/s41467-020-14976-9468025.

[45] Fang H-B, Fang K-T, Kotz S. The meta-elliptical distributions with given marginals. J Multivariate Anal 2002;82(1):1–16. https://doi.org/10.1006/jmva.2001.2017.

[46] Yang E, Ravikumar PK, Allen GI, Liu Z. On poisson graphical models. Adv Neural Inf Process Syst 2013:1718–26.

[47] Karlis D, Meligkotsidou L. Multivariate poisson regression with covariance structure. Stat Comput 2005;15(4):255–65. https://doi.org/10.1007/s11222-005-4069-4.

[48] Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J Roy Stat Soc Ser B (Stat Methodol) 2014;76(2):373–97. https://doi.org/10.1111/rssb.12033.

[49] Allen GI, Liu Z. A local poisson graphical model for inferring networks from sequencing data. IEEE Trans Nanobiosci 2013;12(3):189–98. https://doi.org/10.1109/tnb.2013.2263838.

[50] Irrthum A, Wehenkel L, Geurts P, et al. Inferring regulatory networks from expression data using tree-based methods. PloS One 2010;5(9):. https://doi.org/10.1371/journal.pone.0012776e12776.

[51] Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, Aerts S. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics 2019;35(12):2159–61. https://doi.org/10.1093/bioinformatics/bty916.

[52] Chan TE, Stumpf MP, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst 2017;5 (3):251–67. https://doi.org/10.1016/j.cels.2017.08.014.

[53] Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. Saver: gene expression recovery for single-cell rna sequencing. Nat Methods 2018;15(7):539–42. https://doi.org/10.1038/s41592-018-0033-z.

[54] Mongia A, Sengupta D, Majumdar A. Mcimpute: matrix completion based imputation for single cell rna-seq data. Front Genet 2019;10:9. https://doi.org/10.3389/fgene.2019.00009.

[55] Meinshausen N, Bühlmann P. Stability selection. J Roy Stat Soc Ser B (Stat Methodol) 72 (4):2010;417–473. doi:10.1111/j.1467-9868.2010.00740.x..

[56] Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol 2016;17(1):173. https://doi.org/10.1186/s13059-016-1033-x.

[57] Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10. https://doi.org/10.1093/nar/30.1.207.

[58] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;161(5):1187–201. https://doi.org/10.1016/j.cell.2015.04.044.

[59] Mukherjee S, Carignano A, Seelig G, Lee S-I. Identifying progressive gene network perturbation from single-cell rna-seq data. In: 2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2018. p. 5034–5040. doi:10.1109/EMBC.2018.8513444.

[60] Przybyla LM, Voldman J. Attenuation of extrinsic signaling reveals the importance of matrix remodeling on maintenance of embryonic stem cell self-renewal. Proc Nat Acad Sci 2012;109(3):835–40.

[61] Niwa H, Miyazaki J-I, Smith AG. Quantitative expression of oct-3/4 defines differentiation, dedifferentiation or self-renewal of es cells. Nat Genet 2000;24 (4):372–6. https://doi.org/10.1038/74199.

[62] Wang Y, Fu Y, Yan Z, Zhang X, Pei M. Impact of fibronectin knockout on proliferation and differentiation of human infrapatellar fat pad derived stem cells. Front Bioeng Biotechnol 2019;7:321. https://doi.org/10.3389/fbioe.2019.00321.

[63] Hughes M, Dobric N, Scott IC, Su L, Starovic M, St-Pierre B, Egan SE, Kingdom JC, Cross JC. The hand1, stra13 and gcm1 transcription factors override fgf signaling to promote terminal differentiation of trophoblast stem cells. Dev Biol 2004;271(1):26–37. https://doi.org/10.1016/j.ydbio.2004.03.029.

[64] Bernardoni R, Vivancos V, Giangrande A. glide/gcmis expressed and required in the scavenger cell lineage. Dev Biol 1997;191(1):118–30. https://doi.org/10.1006/dbio.1997.8702.

[65] Zhang H, Nieves JL, Fraser ST, Isern J, Douvaras P, Papatsenko D, D'Souza SL, Lemischka IR, Dyer MA, Baron MH. Expression of podocalyxin separates the hematopoietic and vascular potentials of mouse embryonic stem cell-derived mesoderm. Stem Cells 2014;32(1):191–203. https://doi.org/10.1002/stem.1536.

[66] Aoyama K, Oritani K, Yokota T, Ishikawa J, Nishiura T, Miyake K, Kanakura Y, Tomiyama Y, Kincade PW, Matsuzawa Y. Stromal cell cd9 regulates differentiation of hematopoietic stem/progenitor cells. Blood J Am Soc Hematol 1999;93(8):2586–94. https://doi.org/10.1006/bcmd.1999.0237.

[67] Masui S, Ohtsuka S, Yagi R, Takahashi K, Ko MS, Niwa H. Rex1/zfp42 is dispensable for pluripotency in mouse es cells. BMC Dev Biol 2008;8(1):45. https://doi.org/10.1186/1471-213X-8-45.

[68] van den Boom V, Kooistra SM, Boesjes M, Geverts B, Houtsmuller AB, Monzen K, Komuro I, Essers J, Drenth-Diephuis LJ, Eggen BJ. Utf1 is a chromatin-associated protein involved in es cell differentiation. J Cell Biol 2007;178 (6):913–24. https://doi.org/10.1083/jcb.200702058.